Diatom interactions in the open ocean : from the global patterns to the single cell

Flora Vincent

Université Paris Descartes

Ecole Doctorale : Frontières du Vivant ED 474

Ecology and Evolutionary Biology Section,

Institut de Biologie de l'Ecole Normale Supérieure, CNRS UMR8197, INSERM U1024, Paris

Diatom interactions in the open ocean: from the global patterns to the single cell

Par **Flora VINCENT**

Thèse de doctorat de Biologie

Dirigée par Chris BOWLER

Présentée et soutenue publiquement le 21 Novembre 2016

Devant un jury composé de :

BOWLER, Chris – DRCE	Directeur de thèse
CHAVE, Jérôme - DR2	Rapporteur
LEGRAND, Catherine - Professeure	Rapporteur
THEBAULT, Elisa – CR1	Membre du Jury
LEBLANC, Karine – CR1	Membre du Jury

Résumé (français) :

Les diatomées sont des micro-algues unicellulaires, qui jouent un rôle primordial dans l'eco-système marin. En effet, elles sont responsables de 20% de l'activité photosynthétique sur Terre, et sont à la base de la chaîne alimentaire marine, toujours plus menacée par le changement climatique.

Les diatomées établissent diverses interactions microbiennes avec des organismes issus de l'ensemble de l'arbre du vivant, à travers des méchanismes complexes tels que la symbiose, le parasitisme ou la compétition. L'objectif de ma thèse a été de comprendre comment ces interactions structurent la communauté du plancton, à grande échelle spatiale. Pour ce faire, j'ai développé de nouvelles approches basées sur le jeu de données inédit de *Tara* Océans, une expédition mondiale qui a exploré la diversité et les fonctions des microbes marins, en récoltant plus de 40.000 échantillons à travers 210 sites autour du monde.

Grâce à l'analyse de réseaux de co-occurrence microbiens, je montre d'une part que les diatomées agissent comme des « ségrégateurs répulsifs » à l'échelle globale, en particulier envers les organismes potentiellement dangeureux tels que les prédateurs et les parasites, et d'autre part que la co-occurrence des espèces ne s'explique qu'en minorité par les facteurs environnementaux. Grâce à la richesse des données *Tara* Océans, j'ai par ailleurs permis la charactérisation d'une interaction biotique impliquant une diatomée et un cilié hétérotrophe à l'échelle de l'eco-système, illustrant de surcroît le succès des approches dirigées par les données. Dans l'ensemble, ma thèse contribue à notre compréhension des interactions biotiques impliquant les diatomées, de l'échelle globale à la cellule unique.

Title:
Diatom interactions in the open ocean: from the global patterns to the single cell.

Abstract :
Diatoms are unicellular photosynthetic microeukaryotes that play a critical role in the functioning of marine ecosystems. They are responsible for 20% of global photosynthesis on Earth and lie at the base of marine food webs, ever more threatened by climate change.

Diatoms establish microbial interactions with numerous organisms across the whole tree of life, through complex mechanisms including symbiosis, parasitism and competition. The goal of my thesis was to understand how those biotic interactions structure the planktonic

community at large spatial scales, by using new approaches based on the unprecedented *Tara* Oceans dataset, a unique and worldwide circumnavigation that collected over 40.000 samples across 210 sites to explore the diversity and functions of marine microbes.

Through the analysis of microbial association networks, I show that diatoms act as repulsive segregators in the ocean, in particular towards potentially harmful organisms such as predators as well as parasites, and that species co-occurrence is driven by environmental factors in a minority of cases. By leveraging the singularity of the *Tara* Oceans data, I provide a comprehensive characterization of a prevalent biotic interaction between a diatom and heterotrophic ciliates at large spatial scale, illustrating the success of data-driven research. Overall, my thesis contributes to our understanding of diatom biotic interactions, from the global patterns to the single cell.

Mots clés (français) :
Plancton, diatomées, biodiversité marine, interactions microbiennes, réseaux de corrélation microbiens, structure des communautés, cellule unique, biologie écosystémique

Keywords :
Plankton, diatoms, marine biodiversity, microbial interactions, microbial correlation networks, community structure, single cell, ecosystem biology

[A Yumi, Patrick, Thomas, Elina et David.]

Remerciements

Chris, I will never thank you enough for all the guidance and trust you gave me since I first stepped in your office, back in January 2013. I still very clearly remember my feeling when walking out of my interview: it was *Tara* or nothing, and this was because you knew how to communicate and transmit the beauty and challenges of the *Tara* Oceans project. I will always be grateful for your reactivity as well as the time you took to discuss with me several hours per week despite your overbooked agenda. I have been amazed by your incredible capacity to never, ever, let the pressure you had on your shoulders interfere with your role as my supervisor, and I hope to act the same the day I'm in such a position. I am glad I was able to share with you my ideas, my opinions - about science and life in general - as well as my doubts, that you were able to dissipate (or reinforce…) when necessary. Thank you for the freedom and opportunities you gave me to explore, try, fail, travel, meet, try again, fail again, experience, realize I didn't fail that much, that ultimately lead me today to look back on those three years and have a feeling of intense intellectual stimulation and accomplishment. I'm sure our roads will cross again in the future but today, I hope I have grown into a scientist you can be proud of, even though I imagine I have at some times been difficult to manage.

The *Tara* Oceans consortium deserves obvious recognition for the work that is presented in this thesis. Having the chance to discuss regularly with the best punk scientist ever, Eric Karsenti, is a true gift in the life of any PhD student – actually any scientist - that provides passion and perseverance to follow the path of becoming an accomplished researcher. Lucie and Shruti, you have both been crucial in my advancement as a scientist, through your advice, your help, your support, and your listening. You have never let one of my (numerous) questions unanswered, challenged me when necessary, and let me challenge you in return. This is only the start, I hope, of a life long scientific and human relationship, both already well under way. Gipsi L-M., Karoline F. and Jeroen R., I have learned a lot from you and my stay in Leuven; again, I have literally bombarded you with my inquiries, and you have always responded present. Special credits go to Colomban's team in the far far away Roscoff station for hosting me several months; Sarah R., Nico H., Thibaut P., Seb C., Daniel R. and others I met there. Beyond the science, I remember Diplomatico and philosophical/epic nights. And from this whole nebulous network of fantastic scientists, I need to aknowledge Damien E. (Ah! La Sardaigne!), Daniele I., Fabrice N., Lionel G., Shini S., Emilie V., Julie P., Eric P., John D., Adriana Z., for being a cohort of scientists from which I learned a lot just by looking, listening and exchanging. *Tara* Oceans was literally the best lab in which to do my PhD because yes, *Tara* is a kind of huge immaterial lab with no doors between teams.

I would also like to thank the members of the jury, Jérôme C. and Catherine L. that review this work, which I hope will meet the standards of what you expect accomplished PhDs to deliver, as well as Elisa T. and Karine L. for participating to my defense. Thanks to Elodie V. and Laure G. for sitting in my thesis committee. Special thanks Catherine for giving Emile and I the freedom to organize the Gordon's Social Activities; this opens the door to many thrilling opportunities, such as the being the chair of the next GRS in 2018!

To all the past and present people in my lab, who have all been so supportive over the past years. Leila, Achal, Omer, Richard, Anne-Flore, Zhanru, Heni, Catherine, Martine, Imen, but also Camille and Sandrine. Working at your contact was a daily pleasure, and you all managed to create a true atmosphere where working is a thrill. To my open office mates, Magali [aaaa tes gateaux], Jérôme [aaaa ton café], Ana [for putting me back in the right track when I was a bit down], Leandro [for being my mentor on what why when who where in science and being the only one to go to weird physics seminars with me], and recently Fabio and Fede, with whom I'm sure the coming months will be really fun. To my lunch mates from the Navarro's team, I have appreciated every single debate we initiated just for the pleasure of arguing, from feminism to ethics in science, the impact of high technologies and life in general. Tu croyais que j'allais t'oublier, toi, le premier à m'adresser la parole le jour où j'ai débarqué à la fête de Noël de *Tara* ? Yann, tu as un coeur tellement grand qu'il rentrerait pas dans ton futur monospace de famille nombreuse. J'ai peu de mots pour exprimer la reconnaissance que j'ai pour toi, ton calme, ta patience, ta gentillesse et ta compétence.

Merci bien sûr le CRI. Pascal, François, Ariel (et tant d'autres !); vous avez créé une formation qui est la meilleure chose qui me soit arrivée de toute ma scolarité. Merci à vous de rappeler ce qu'est vraiment la science; une aventure humaine, et une persévérance avant tout animée par l'envie de savoir, de comprendre, de questionner, de transmettre, d'essayer. Que les frontières, quelles qu'elles soient, sont celles que nous nous fixons nous-mêmes. Vous donnez les clefs à vos étudiant-e-s et les laissez faire; mieux que ça, vous les écoutez pour aller encore plus loin. Et oui, on en a des choses à dire (à Jean et Thomas pour nos débats sur la science). Car évidemment, qui dit CRI, dit WAX Science. WAX, école de la vie et de l'audace, source de tant de situations cocaces. Des couloirs de l'Assemblée Nationale au forum étudiant-e-s d'Amiens, de Mexico à Saint Nazaire, de la scène des Frigos à celle de Chambord, des papys d'IESF aux ados des Solidays, d'une idée folle à un vrai projet de vie. La relêve est là, c'est peut-être la plus belle "exit strategy" qu'on pouvait espérer, et mes années de thèse resteront profondément marquées par cette aventure collective rendue possible grâce à vous tous, les bénévoles qui ont oeuvré sans relâche pour promouvoir cette science belle, créative, utile et surtout intense. Vu que les remerciements ne doivent pas être un chapitre entier, je n'en nommerai qu'une. Aude, tu es exceptionnelle.

Et parce que la thèse et la science ne font pas tout, merci aux potes sans qui la vie ne serait rien. Merci Didier, qui aura insufflé en moi l'amour de la mer, qui m'a offert mes premières plaquettes plastifiées d'identification de la faune et la flore méditerranéenne, et a sonné le début d'une longue route vers l'exploration de nos fonds marins, qui rythme aujourd'hui mon quotidien. Et puis les ancien-ne-s, valeurs sûres dans ces temps qui bougent: Paul, Noémie, Louis, M.M.M, Elsa, Julie, Guillemette, Etienne, Adrien, d'avoir toujours été là. Les musiciens de Salon Alpha et les Agros, trop nombreux pour être nommé-e-s: Victor, Chloé, Marina, Ed, Thib, et les plongeurs très présent-e-s ces derniers mois. Au Judo Club Sorbier, à qui j'aurais au moins appris la différence entre le zooplankton et le phytoplankton, entre deux virées folles à Chambéry et un tatami. Raymond, Thabo, Camille, Karl, Ulysse, Magali, Julie, Milan, Xav, Clarisse [rencontrer Cédric, meilleure chose arrivée à WAX, et grâce à toi], Nico, Habib, Lulu, Laeti, Dams et tous les autres…Vous avez rythmé mes semaines au son des Hajimé et des verres qui trinquent, apporté l'équilibre nécessaire à l'esprit sain dans un corps sain (enfin…). Celà a été trois années de bonheur et de partages précieux avec vous.

J'ai l'impression d'avoir accompli quelque chose en faisant une thèse de trois ans, mais qu'en est-il de vous, mes parents, qui avez passé plus de 15 ans à l'oeuvre? Ces longs remerciements ne sont que le reflet du temps que vous avez pris pour faire de nous trois des enfants chanceux. Chanceux d'étudier, de voyager, d'être curieux, de faire ce que nous aimons, d'avoir le choix, de s'en donner les moyens, d'être tournés vers l'avenir, de vous avoir. Alors à l'ensemble de ma famille qui s'aggrandit, mes grands parents, Véro et Christian, et surtout mes parents, mon frère et ma soeur, je veux vous dire merci, mais je sais que vous serez là encore demain, après demain et plus loin encore.

Et enfin, à David, pour m'avoir accompagnée 5 années et demi de ma vie, en souvenir de tous les moments profondément heureux que nous avons vécus.

Table of contents

Table of figures

Chapter 1: Introduction

Summary

1.1. Marine planktonic ecosystems

1.1.1. Odyssey of plankton research

The oceans comprise the largest continuous ecosystem on Earth, and 98% of its biomass is composed of organisms that are invisible to the naked eye: marine microbes, many of which live as "plankton". The word plankton comes from the Greek *planktos* "errant"; it designates organisms that live in the water column and are unable to swim against the current (Lalli et al., 1993). Marine plankton is therefore composed of bacteria, protists (unicellular eukaryotes), fungi, viruses, archaea, but also of eggs and larval stages of larger animals.

While the larger sized plankton have been studied for more than a century, it is only with the advent of modern techniques since the late 1970's that the abundance of microbes and viruses has been appreciated (Hobbie et al., 1977). It was shown since then that a liter of seawater contains up to $10^{\wedge 8}$ bacteria (Whitman et al., 1998). In the 1990's, the first cultivation-independent assessment of marine bacterial diversity - through rRNA analysis - showed that they were highly diverse, and that most groups in the ocean were previously unknown (Stahl et al., 1984). Simultaneously, the discovery of viruses in the ocean, reaching nearly $10^{\wedge 10}$ particles per liter (Wilhelm et al., 2008), added a new layer of complexity in our understanding of what generates and maintains microbial diversity (Bergh et al., 1989). Diversity of protists - unicellular eukaryotes - was coined in the early 2000 using rRNA, making our knowledge of the majority of planktonic diversity fairly recent (Moon-van der Staay et al., 2001). Exploration of global patterns of marine microbial communities and diversity became quantitative with pyrotag sequencing in the mid 2000's (Sogin et al., 2006), opening the way for large spatial scale campaigns of marine metagenomic surveys such as the Sorcerer II Global Ocean Sampling survey (Rusch et al., 2007) and the Malaspina deep sea expedition in 2011 (See Annexe A for details on the contribution of genomics to microbial studies). However, our knowledge today is partially restrained by the non-cultivability of 90% to 99% of marine microorganisms in current laboratory settings.

Marine plankton distribution is strongly dependent on abiotic factors such as light and nutrients, turbulence, temperature, salinity, redox potential or pH, as well as biotic factors such as the presence of other planktonic organisms. If local abundance of plankton varies

horizontally, vertically, and seasonally, planktonic organisms are virtually present everywhere in the water column across the ocean.

1.1.2. Importance in Earth's biogeochemical cycles

The metabolism of marine microorganisms composing plankton maintains major biogeochemical cycles on Earth (Falkowski et al., 2008), including that of carbon, oxygen, nitrogen, phosphorus and sulphur. All these chemical elements circulate through the biological and physical world, thanks to a global recycling where chemical compounds are passed from one organism to another, and from one part of the biosphere to another. A simplified version of how plankton contribute to major nutrient cycles illustrates both the complexity and the inter-connectedness of these processes.

Carbon is critical for life because it is the building block of all organic compounds. Carbon dioxide (CO_2) is exchanged between the surface ocean and the atmosphere, but must be transformed – or "fixed" - into a usable organic form for living organisms. In the sunlit part of the ocean, CO_2 is fixed mainly through photosynthesis performed by phytoplankton - autotrophic prokaryotes and eukaryotic algae - resulting in the production of oxygen from water (Field, 1998). CO_2 fixation by phytoplankton leads to the creation of dissolved organic carbon (DOC) - any organic matter that is smaller than 0.45 micron - and particulate organic carbon (POC) - organic matter bigger than 0.45 micron - both of which can be respired back to CO_2 by other microorganisms such as zooplankton, the heterotrophic metazoans and protozoans. Part of this organic matter will sink to the ocean floor through a process called the biological pump. What is not mineralized and recycled into carbon dioxide by bacteria during the descent can be buried and stored over geological timescales, eventually leading to the formation of natural gas, petroleum, and kerogen (**Figure 1.1**).

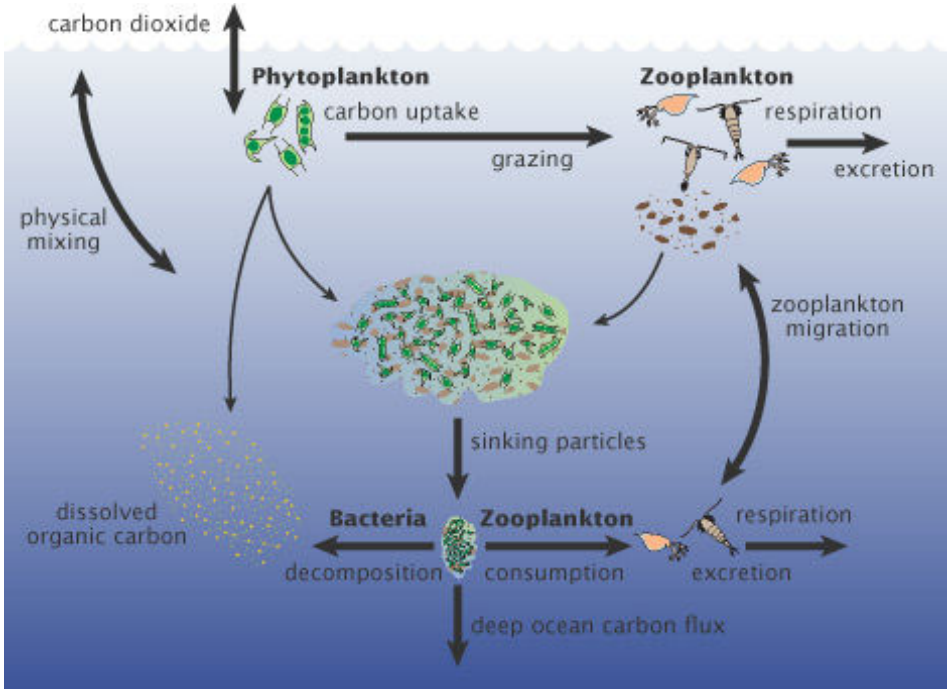


Figure 1.1. Marine microbes and the carbon cycle in the ocean.
Carbon can be transferred from the atmosphere to the ocean depths through the carbon cycle, in which marine microbes play a key role (U.S. JGOFS).

Nitrogen is one of the most important elements because it is required for synthesis of the basic building blocks of life; many organisms depend on ammonium (NH_4^+), nitrate (NO_3^-), or dissolved organic nitrogen (DON) to synthesize amino acids, nucleic acids or cell walls. The most abundant form of nitrogen on Earth is atmospheric dinitrogen (N_2), but in the ocean only a few specialized bacteria and archaea - equipped with a nitrogenase enzyme - are capable of fixing gaseous nitrogen into a bioavailable compound such as ammonium (Galloway et al., 2013). Ammonium can be rapidly assimilated as a nutrient by microorganisms, and be incorporated into living cells. Other microbes extract energy from nitrogen by oxidizing ammonium into nitrite and nitrates through nitrification, ultimately releasing nitrates in the ocean. Some organisms can uptake nitrates, convert them back to nitrogen gas via denitrification, essentially putting nitrogen back into the atmosphere. Incomplete denitrification may lead to the formation of nitrous oxide (N_2O), which is a strong greenhouse gas (**Figure 1.2**).

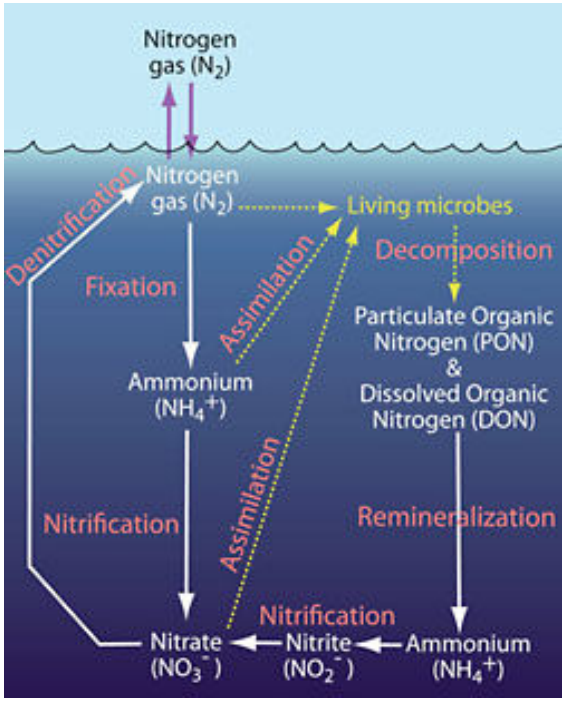


Figure 1.2. Marine microbes and the nitrogen cycle in the ocean.
The nitrogen cycle is the biogeochemical cycle by which nitrogen is converted into various chemical forms as it circulates among the atmosphere and terrestrial and marine ecosystems (CMORE).

1.1.3. Connected micro-organisms: the example of the food web

The importance of plankton in our ecosystem is not only embodied by a vast grid of exchanged metabolic currencies. Plankton play a crucial role in the very structure of the species network in the ocean, illustrated by the role of plankton in the marine food web. Indeed, beyond the necessary cataloguing of which species are present, how they interact complements our understanding of ecosystem structure and functioning.

Plankton are the foundation of **classical marine food webs** by occupying a critical bottom trophic level. Phytoplankton live in the surface water where they access enough light to perform photosynthesis and convert inorganic carbon dissolved in the surface waters to organic carbon (Field et al., 1998). Phytoplankton are then grazed upon by zooplankton, serving as a crucial link between primary producers and the rest of the marine food web, including higher consumers such as large fish, and ultimately, top predators. However the

view that fixed carbon is transferred vertically up trophic food webs has been challenged by the discovery of abundant heterotrophic prokaryotes, revealing that a large proportion of the flux of matter and energy in the marine food web passes through these smaller organisms in the form of dissolved organic matter; a concept named the microbial loop .

The microbial loop is the process by which dissolved organic material (DOM) derived from primary producers such as phytoplankton is foraged and recycled by heterotrophic bacteria that use DOM as a food source (Hobbie et al., 1972; Azam et al., 1983). These bacteria are consumed by protozoans, then fed upon by zooplankton, and the link to the traditional food web is established. Marine bacteria hold a significant influence in this loop, as they allow for an energy pathway that may have otherwise been lost (**Figure 1.3**). Viruses, by infecting and lysing marine bacteria and phytoplankton, are also vital players in the fixation and cycling of key elements, such as carbon, nitrogen, and phosphorus. Through the **viral shunt**, viral lysis of microbial cells releases dissolved organic matter and particulate organic matter back into the microbial loop. By doing so, viruses drive carbon and nutrients of these cells away from grazers, and redirects it to other microorganisms through the form of DOM. It has been estimated that ocean viruses might turn over as much as 150 gigatons of carbon per year (Suttle, 2007).

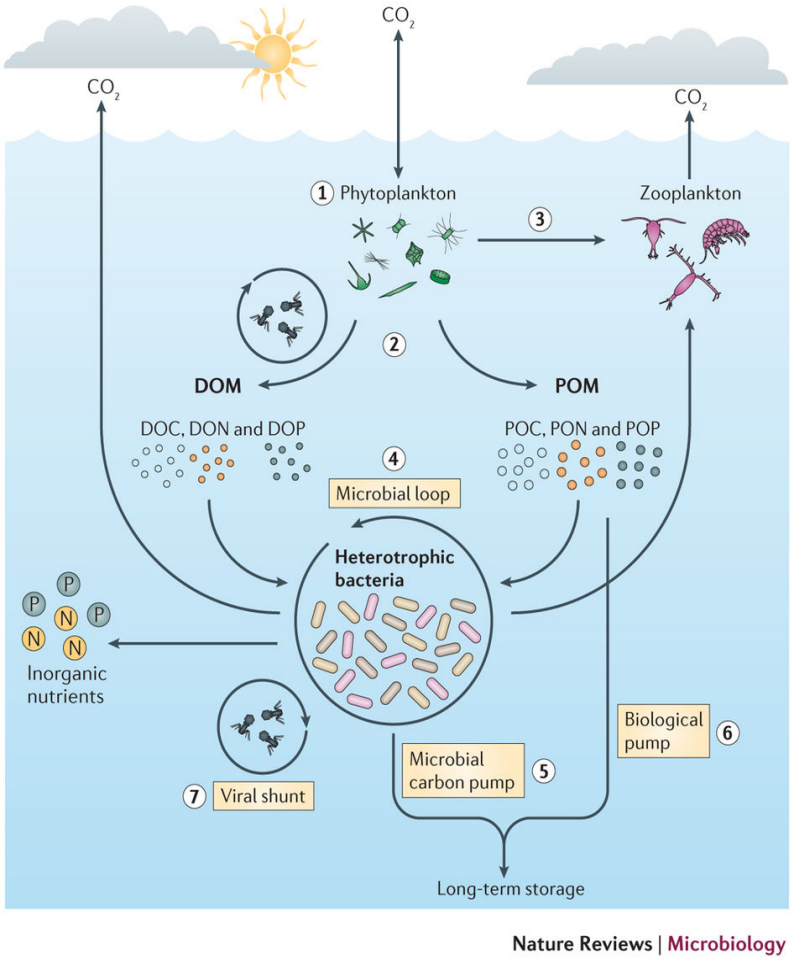


Figure 1.3. Carbon cycle processes mediated by micro-organisms.
Key processes of the marine carbon cycle include the conversion of inorganic carbon (ex: CO_2) to organic carbon by phytoplankton (step 1); the release of dissolved organic matter, dissolved organic phosphorous and particulate organic matter by phytoplankton (step 2); the consumption of phytoplankton biomass by zooplankton (step 3) and the mineralization (release of CO_2 via respiration) and recycling of organic matter by diverse heterotrophic bacteria (the microbial loop; step 4). Some heterotrophic bacteria are consumed by zooplankton, and the carbon is transferred up the food web. Heterotrophic bacteria also contribute to the remineralization of organic nutrients, to inorganic forms, which are then available for use by phytoplankton. The microbial carbon pump (step 5) refers to the transformation of organic carbon into recalcitrant dissolved organic carbon that resists further degradation and is sequestered in the ocean for thousands of years. The biological pump (step 6) refers to the export of phytoplankton-derived particulate organic matter from the surface oceans to deeper depths via sinking. Finally, the viral shunt (step 7) describes the contributions of viral-mediated cell lysis to the release of dissolved and particulate matter from both the phytoplankton and bacterial pools (Buchan et al., 2014).

Biotic interactions such as viral infection or other types of grazing by prokaryotes have considerably complemented our vision of the marine food web, and thus of the related carbon cycle. Many other biotic interactions are likely to impact ecology and ecosystem functioning of the ocean hence the critical importance of plankton ecology in the Earth

system can hardly be overstated. Yet the study of marine microbial ecology as a major component of marine science has a fairly short history (Pomeroy, 1974; Azam et al., 1983), in particular that of planktonic micro-eukaryotes also named protists.

1.1.4. The forgotten marine micro-eukaryotes: protists

The study of diversity and function of marine microbial communities has largely focused on bacteria in the past decades, and to a lesser extent on archaea and viruses. Paradoxically, much lesser attention has been given to single-celled eukaryotic microbes - the **protists** – despite them being some of the first microbial taxa observed, that also play a key ecological role as primary producers, consumers, decomposers and trophic links that are extremely abundant with $10^{\wedge 6}$ eukaryotic cells per liter of seawater (Brown et al., 2009). Reasons for this dismissal are multiple: their misleading first naming as “animalcules” (van Leewenhoek) could have semantically related them to larger multicellular organisms thus excluding them from subsequent research, that has focused largely on prokaryotic microbes; the unfair granting of bacterial mortality to viruses in the ocean, to the expense of protistan bacterivory research; the limitations in sequencing and computation techniques to deal with protistan large genome sizes and complexity (Caron et al., 2009).

Nonetheless, protists are present basically in every branch of the eukaryotic tree of life and comprise the bulk of eukaryotic phylogenetic diversity (**Figure 1.4**). When we picture the divergence between multicellular eukaryotes (animals and fungi) just in the single supergroup of Opisthokonta, we come to appreciate the breadth of how protists populate the entire Eukarya domain. From free-living amoeboid taxa (Amoebozoa) and parasites (Excavates) to strontium sulfate or calcium carbonate skeleton builders (Rhizaria), protists display countless forms, sizes, and trophic activities. Understanding protistan ecology and evolution is likely to bring insights into the deep branches of the eukaryotic tree of life, and thus on the origins of multicellular taxa (Baldauf, 2003).

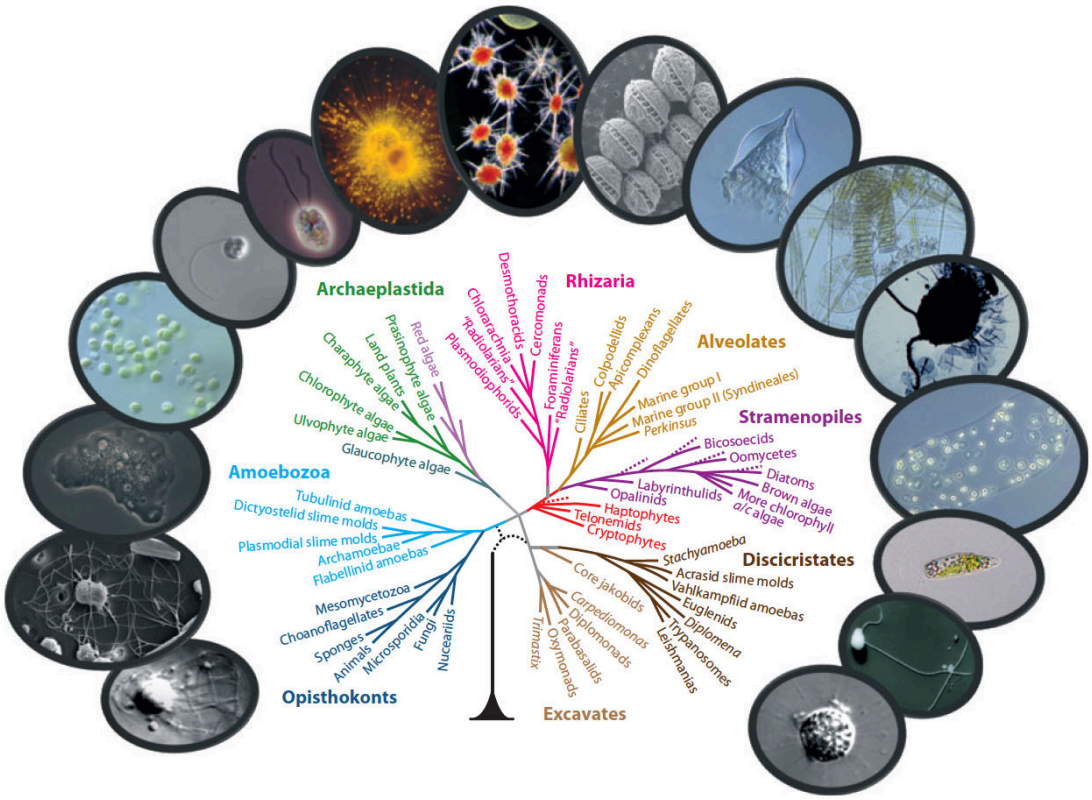


Figure 1.4. The eukaryotic tree of life populated with protists.
Phylogenetic breadth among protists. Single-celled species (protists) occur in every supergroup within the domain Eukarya, and constitute the entirety of a number of them (Caron, 2011).

In order to increase our knowledge about marine protists, I decided to investigate the impact of biotic interactions on community structure at large spatial scales, focusing my work on a key phytoplankton group: diatoms.

1.2. Diatoms, pivotal in the plankton community

1.2.1. General biology of diatoms

1.2.1.1. From art to science

Why diatoms? Diatoms (Bacillariophyta) are unicellular photosynthetic microalgae that display a unique and distinctive feature that sparked my curiosity for it was so fascinating and atypical: they are enveloped in a silica cell wall, called the frustule, that comes in many shapes and has inspired Victorian artists in the late 19th century, a period during which arts and science were intertwined. By collecting diverse diatoms in aquatic environments,

artists/scientists would meticulously shape them under the form of “diatom arrangements” leading to a hypnotizing art piece (**Figure 1.5**).

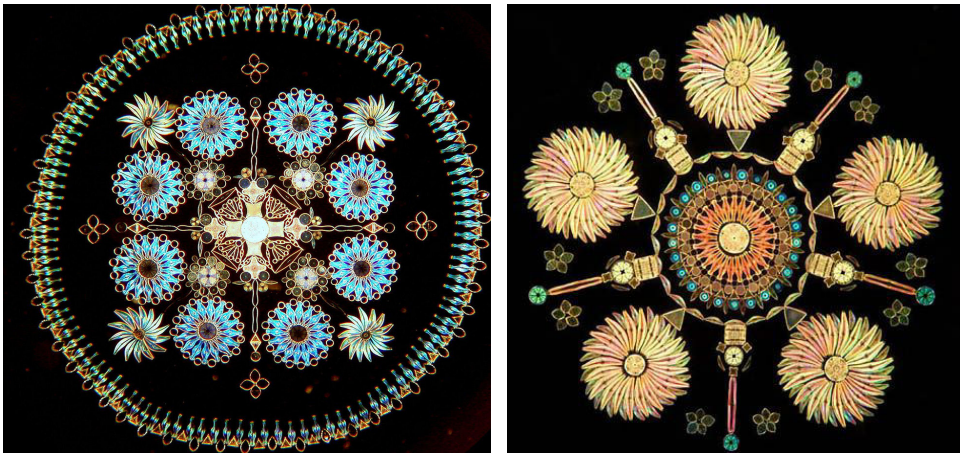


Figure 1.5. Diatom arrangements.
(Source: Google Image).

The first illustration of a diatom goes back to 1703; it was made by an unknown Englishman and published in the Philosophical Transactions of the Royal Society of London. The drawing is a representation of the freshwater diatom *Tabellaria* sp., that forms colonies of cells connected at their corners (**Figure 1.6**).

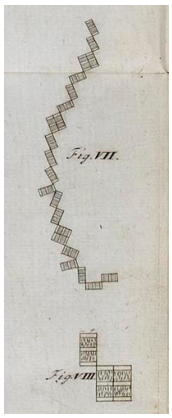


Figure 1.6. First illustration of the diatom *Tabellaria* (Anonymous, 1703).

Advances in electron microscopy during the late 1950s opened avenues for discoveries in the field of diatom research, allowing detailed examination of diatom morphological diversity, found in freshwater, sea ice, soils, or the open ocean. However, interest in diatoms goes way beyond their morphological eccentricities and is as wide as their habitat.

In applied ecology, they are used to monitor change in aquatic systems particularly in rivers, as they are sensitive indicators of the ecosystem's health (Round et al., 1990). They are considered as serious candidates for biofuel production as they accumulate high amounts of lipids (D'Ippolito et al., 2015), but are also known to generate a number of interesting biomolecules that have human health benefits or commercial applications, such as omega-3 fatty acids (Petrie et al., 2010). Last but not least, in the marine environment they serve as the basis of the marine food web, supporting the growth of higher organisms. They are significant players in global biogeochemical cycles, as they account for 20% of global primary production (Nelson et al., 1995; Falkowski, 2002) and fix as much carbon as all the tropical rainforests combined, equivalent to ~ 20 Pg of fixed carbon per year (Field, 1998; Mann, 1999).

1.2.1.2. The silica cell wall

The word diatom comes from the Greek *diatomos*, meaning "cut in half", as the silica cell wall is often bilaterally symmetrical (**Figure 1.7**). The diatom frustule is composed of two overlapping thecae: the larger one, named *epitheca*, and the smaller one, the *hypotheca*. Each theca consists of a valve, and a series of girdle bands (**Figure 1.8**)**.** The frustule is nano-patterned with rows of pores, called interstriae, and ribs between them called striae. The pores in the interstriae serve principally for the uptake of nutrients, and exudation of metabolites. The girdle bands are generally simple; however, valves can be much more elaborated, with a flat area called the valve face, and a rim, called the mantle **(Figure 1.9**).

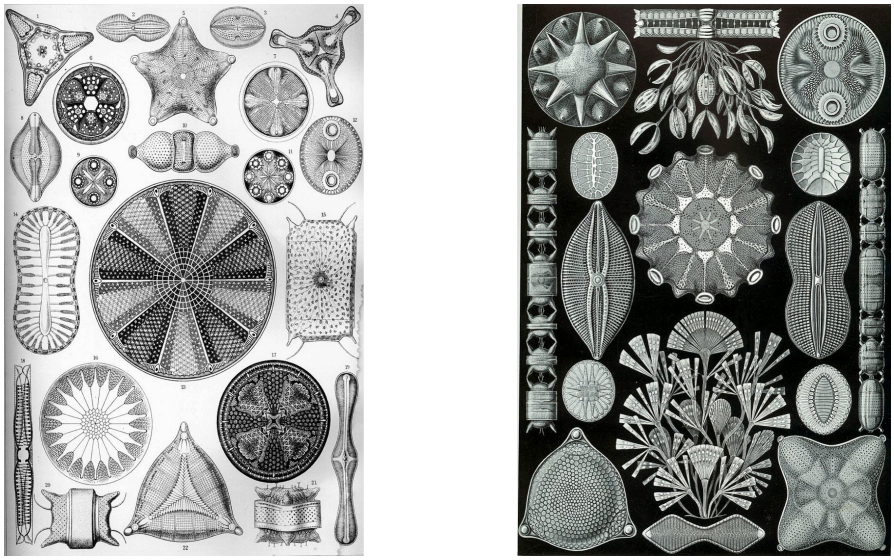


Figure 1.7. Diatom illustrations.
(Haeckel, 1899)

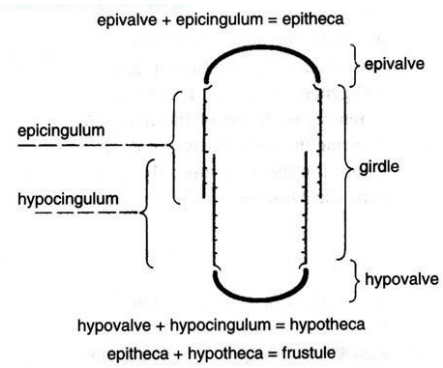


Figure 1.8. Diatom thecae.
(Benten and Harper, 2013)

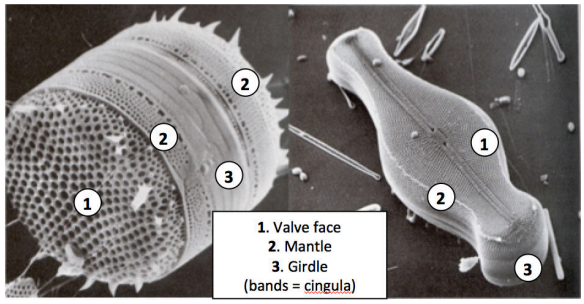


Figure 1.9. Diatom valve, mantle and girdle.
(Photo: Yuki Sawai)

The frustule elements of diatoms are formed in silica deposition vesicles (SDVs) within the cell under mildly acidic conditions. Calcium binding glycoproteins, called frustulins, coat the frustule and are then exocytosed from the protoplast (Vrieling et al., 1999). The mechanisms by which diatoms execute their morphogenetic program to biomineralize SiO_2 and achieve such a high diversity of nano and micro patterns is nicely reviewed in Kröger, 2008. This paper also highlights how the discovery of fundamental principles in the molecular mechanism of diatom silica formation has encouraged biomimetic methods, leading to innovations in the field of nanomaterial sciences and technology. Based on the morphological characteristics of valves, including symmetry, morphology and ornamentation, as well as interstriae/striae organisation, two main categories of diatoms are distinguished: the centric and the pennate diatoms (Van den Hoek et al., 1995).

Centric diatoms are composed of two groups: radial centrics, and multipolar centrics. Radial centrics have circular valves and can form chains using tubes called strutted processes, through which chitin filaments are exuded and link them to the adjacent cell. Multipolar centric diatoms can form elongated, triangular, or starlike profiles (i.e., the cell form exhibits polarity of shape (Piganeau, 2012)) even though they also exhibit radial pore organization (**Figure 1.10**).

Pennate diatoms are elongated, and are distinguished by the midrib, from which striae and interstriae extend perpendicularly. Like the centrics, pennate diatoms can be divided in two main groups: the raphid pennates and the araphid pennates. Raphid pennates possess a slit, which enables them to move through a system of traction. Most raphid pennate diatoms are benthic, but a few have developed a truly marine planktonic lifestyle such as *Fragilariopsis* and *Pseudo-nitzschia* (Falkowski, 2004). *Pseudo-nitzschia* is a planktonic representative, which evolved from a benthic ancestry (Piganeau, 2012), and acquired the capacity to form chains by using its raphe and positioning cells for them to be attached by their valve apices. On the contrary, most araphid pennates are benthic, although a few genera such as *Lioloma*, *Thalassionema*, or *Asterionella* are planktonic (**Figure 1.11**)**.**

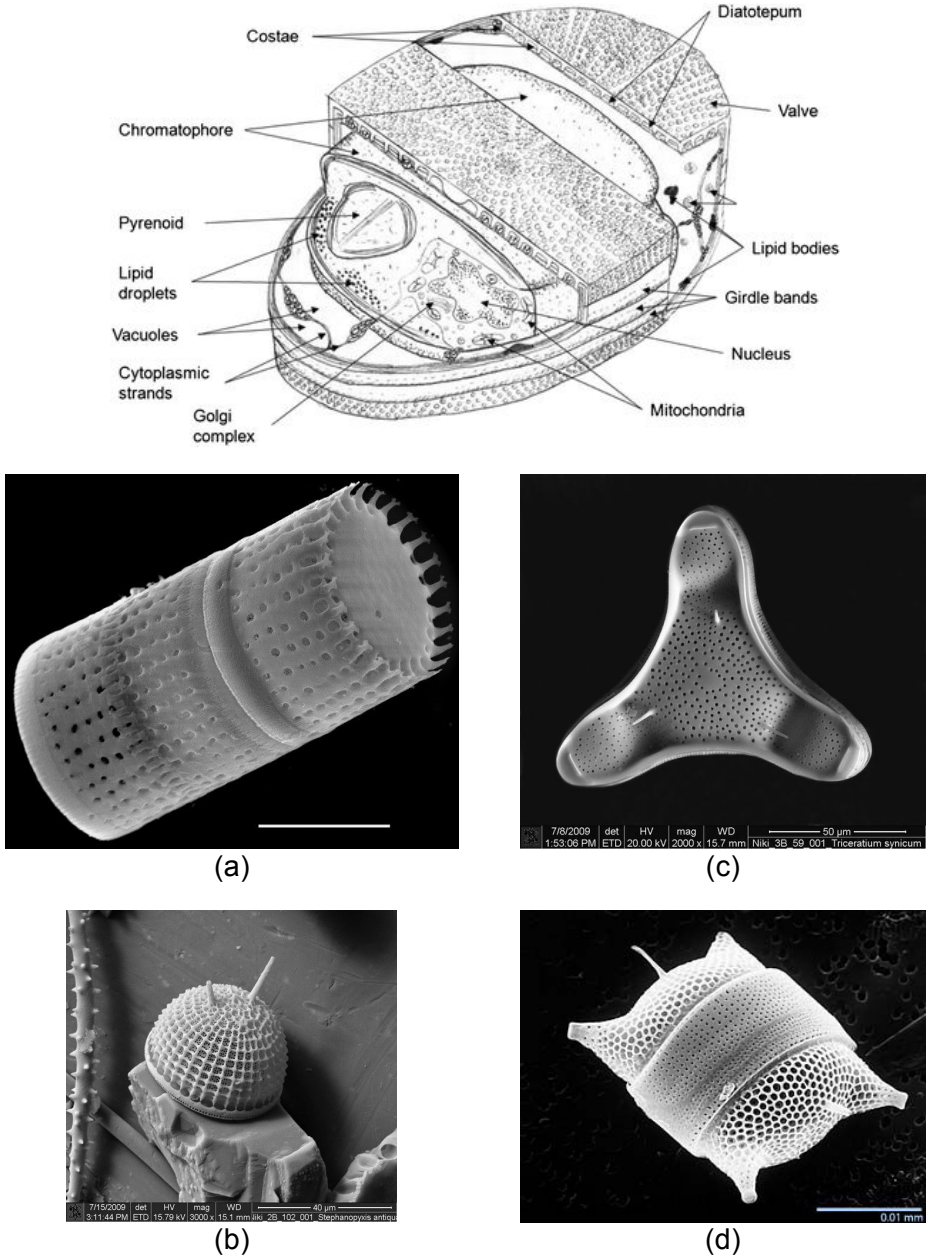


Figure 1.10. Centric diatom structure under electron microscopy.
(a) *Aulacoseira lirata*, Ehrenberg – radial (Photo: Hartley, 1986)
(b) *Stephanopyxis antiqua* - radial (Photo: Ekaterina Nikitina)
(c) *Triceratium synicum* – multipolar (Photo: Ekaterina Nikitina)
(d) *Odontella sp.* – multipolar (Photo: PhytoPedia)

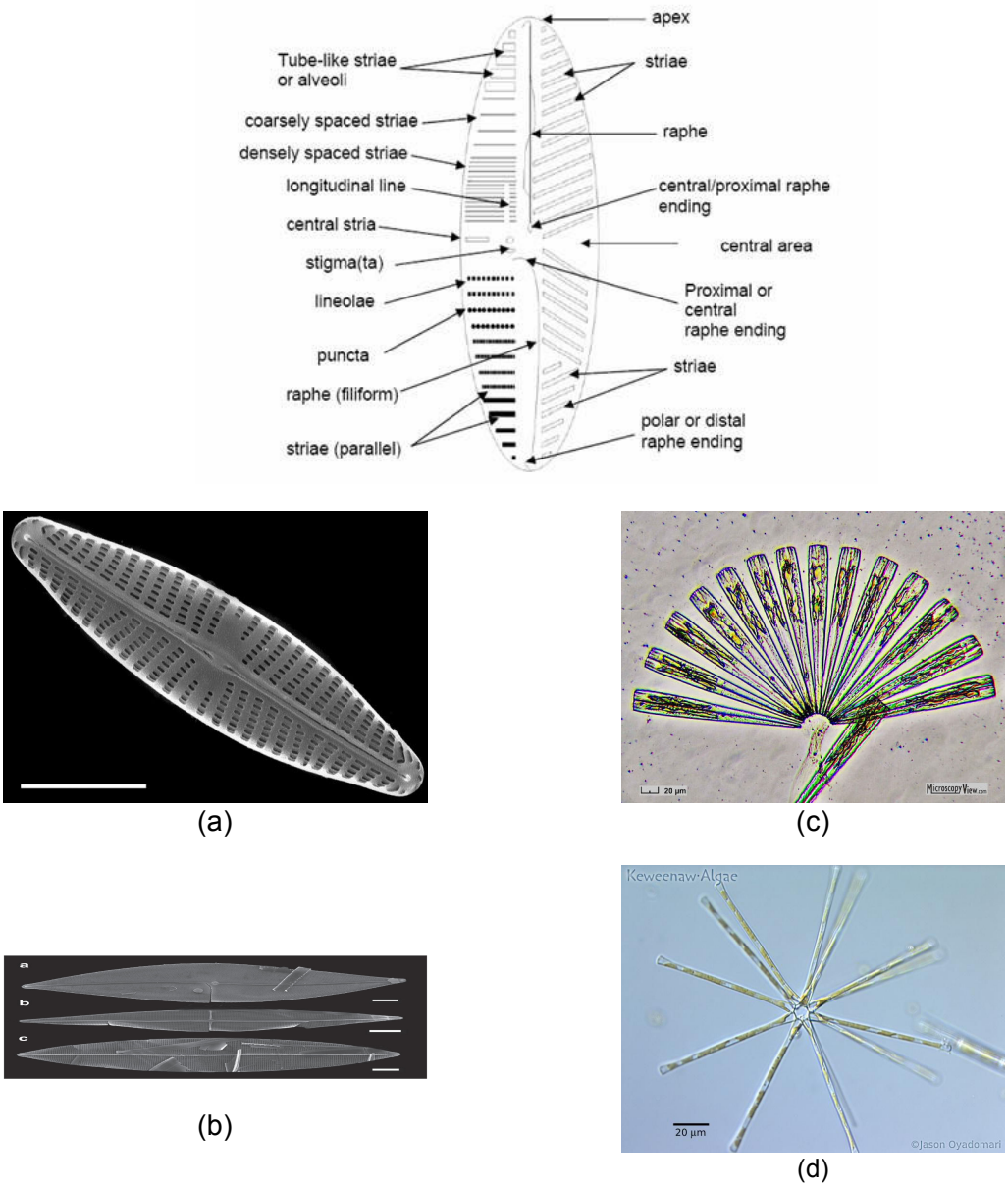


Figure 1.11. Pennate diatom structure.
(a) *Navicula veneta* – raphid (scale bar = 5 micron) (Photo: westerndiatoms.colorado.edu)
(b) *Pleurosigma stuxbergii var. Rhomboides* - raphid (Scale bars, 10 µm) (Photo: Brown et al., 2014)
(c) *Licmophora splendida* – araphid (Photo: Robert Lavigne)
(d) *Asterionella formosa* – araphid (Photo : Jason Oyadomari)

1.2.1.3. Diatom sexual cycle

Diatoms have a diplontic life cycle: the vegetative cells are diploid, and the only haploid cells are the gametes. They are able to reproduce both sexually and asexually, but duplicate primarily by a unique “shrinking division” mode of asexual reproduction. During clonal cell

division, the two valves get separated, each of them forming the epivalve of the daughter cells. New hypovalves are secreted within the daughter cells. The epivalve of the mother cell becomes the epivalve of one daughter cell, hence conserving the same size. On the contrary, as the hypovalve of the mother cell becomes the epivalve of the second daughter cell, the daughter cell has to build a smaller hypovalve, leading to a diminishing average cell diameter with successive mitotic divisions (**Figure 1.12**).

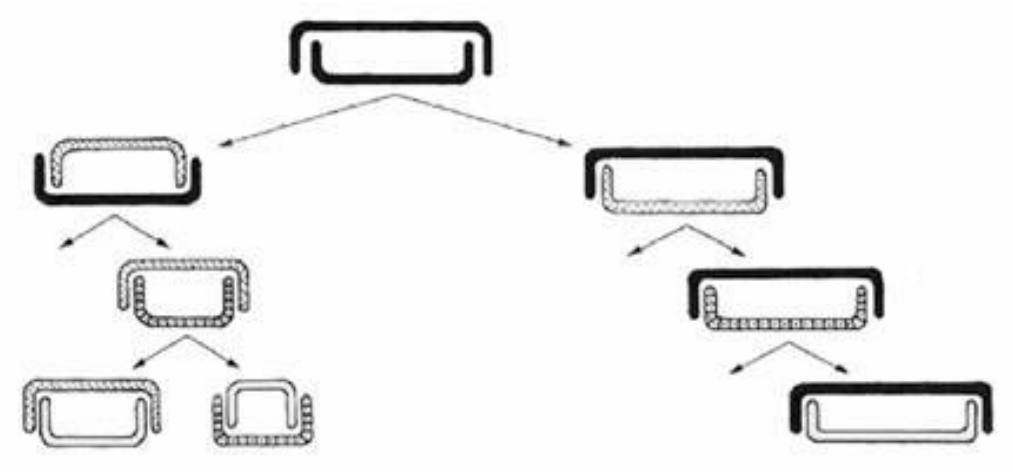


Figure 1.12. Diatom size reduction : the MacDonald-Pfitzer hypothesis.
At each division the new valve is always formed within the parental theca, causing the average size of the frustules in a population to slowly decrease as shown down the left arrow (MacDonald 1869; Pfitzer 1869).

Recovery of the initial cell size is achieved by means of sexual reproduction (Edlund & Stoermer, 1997). Sex includes gamete formation, conjugation, and zygote formation by isometric (for radial centrics and Thalassiosirales) or anisometric (for multipolar centrics and pennates) swelling of zygotes into an auxospore, and the formation of the initial cell structure (**Figure 1.13**).

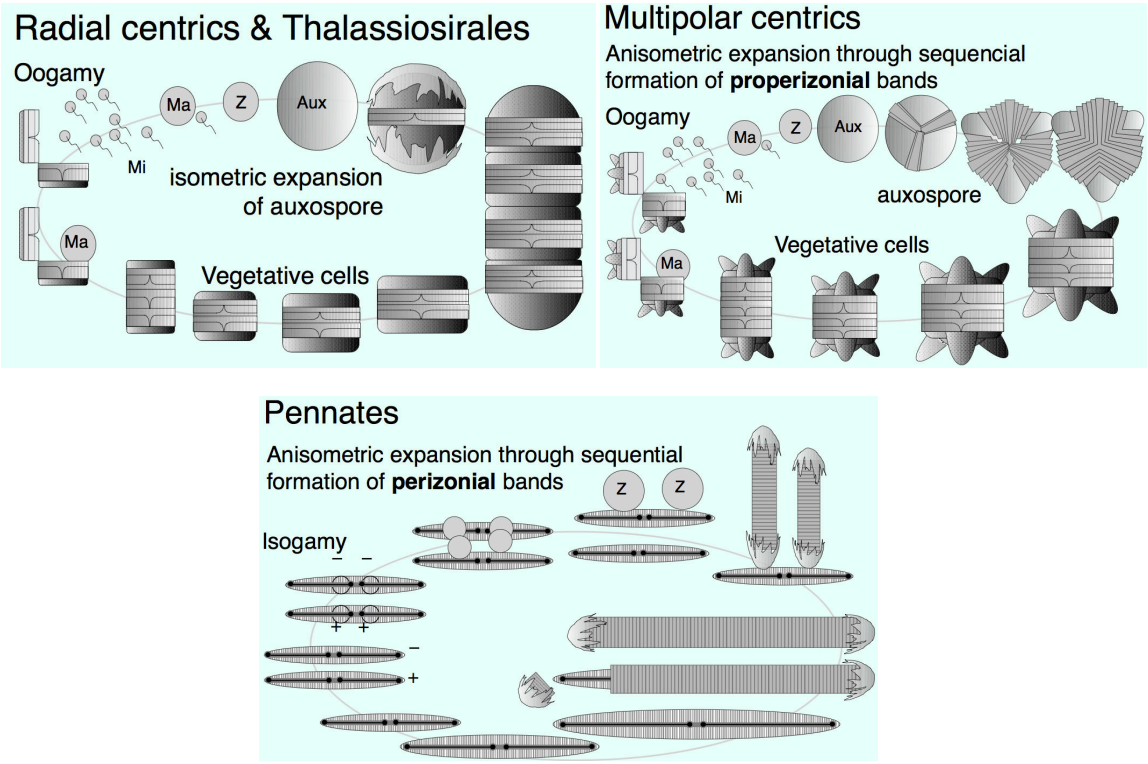


Figure 1.13. Diatom life cycle.
Adapted from (Kooistra et al., 2007) and (Round et al., 1990).

1.2.1.4. Evolutionary history and divergence

Diatoms are an important component of "phytoplankton", a name first used in 1897 that defines a polyphyletic group of single celled organisms that drift with the currents, both in marine and freshwater environments. A trio of major eukaryotic phytoplankton, composed of dinoflagellates, coccolithophores, and diatoms, has dominated the ocean since the Mesozoic Era (251 to 65 million years ago). Phytoplankton represent 1% of the Earth's photosynthesis biomass, yet are responsible for 45% of the Earth's annual net primary productivity (Behrenfeld et al., 1997; Falkowski, 1998). Diatoms alone contribute to 20-25% of global net primary production (Nelson et al., 1995), which is more than all the world's rainforests (Field, 1998). How did diatoms evolve to become such an important player in the global ocean?

Endosymbiotic theory explains the origins of eukaryotic organelles such as mitochondria and chloroplasts (Sagan-Margulis, 1967). Primary endosymbiosis involves the engulfment of

a bacterium by another free-living organism. Early in life history, a primary endosymbiotic event occurred, during which a large, heterotrophic, eukaryotic cell engulfed a small autotrophic bacterium – probably an alphaproteobacteria - that evolved to become the mitochondrion ~1,200 Mya (Shih et al., 2013). A similar process was proposed for the origin of chloroplasts: a eukaryotic host cell, already containing a mitochondrion and a nucleus, is thought to have engulfed - or been invaded by - a cyanobacterium that became the chloroplast, leading to the first oxygenic photosynthetic eukaryotes. Three major lineages arose from this event: the green algae, that use chlorophyll *b* as a light harvesting antenna in photosystem II, the red algae, that use chlorophyll *c* and derivatives as accessory photosynthetic pigments, and finally the glaucophytes (Falkowski, 2004). Secondary endosymbiosis occurs when the product of primary endosymbiosis is itself engulfed by another free-living organism. A few hundred million years after the primary endosymbiosis leading to chloroplasts, a secondary endosymbiosis took place, in which a non-photosynthetic eukaryote acquired a chloroplast by engulfing a red algae leading to Stramenopiles, the superphylum containing diatoms (Keeling, 2004). These successive events were accompanied by massive gene transfers from the swallowed cell to the host (Armbrust, 2004). However, evidence for genes of green algal origins in diatom has challenged this view (Moustafa et al., 2009) suggesting the potential additional engulfment of green algae that led to Chromalveolates – a eukaryotic supergroup that contains diatoms - during secondary endosymbiosis (**Figure 1.14**).

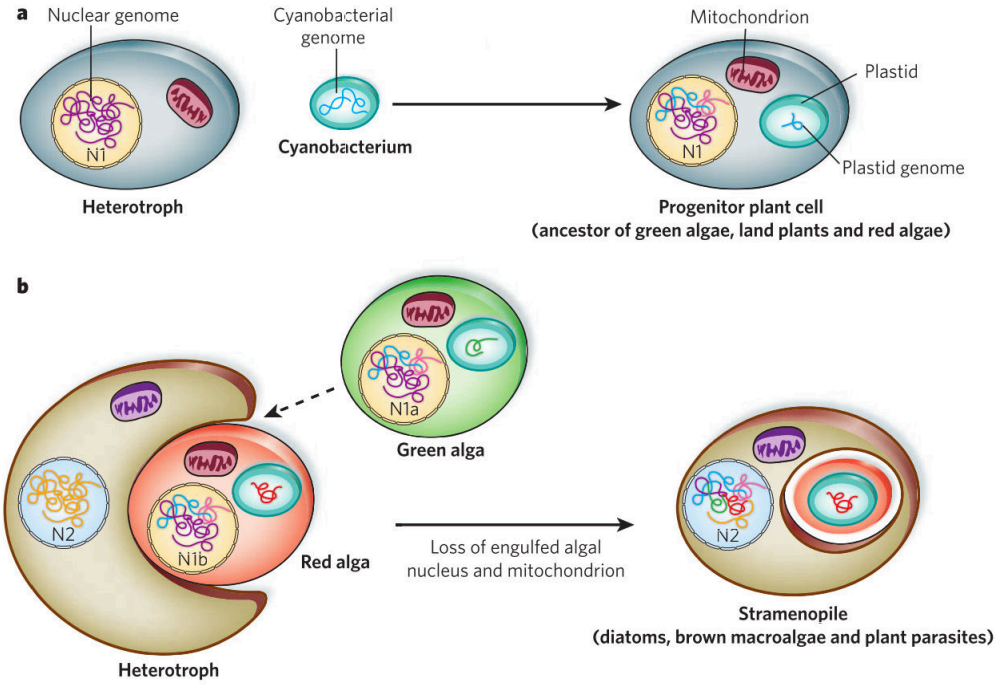


Figure 1.14. Secondary endosymbionsis.
Representation of the origin of diatom plastids through sequential primary (a) and secondary (b) endosymbioses, and their potential effects on genome evolution. a| During primary endosymbiosis, a large proportion of the engulfed cyanobacterial genome is transferred to the host nucleus (N1), with few of the original genes retained within the plastid genome. The progenitor plant cell subsequently diverged into red and green algae and land plants, readily distinguished by their plastid genomes. b| During secondary endosymbiosis, a different heterotroph engulfs a eukaryotic red alga. Potential engulfment of a green algal cell as well is indicated with a dashed arrow. The algal mitochondrion and nucleus are lost, and crucial algal nuclear and plastid genes (indicated in blue, purple and pink) are transferred to the heterotrophic host nucleus (Armbrust, 2009).

The subsequent gains (and losses) of specific genes, largely from bacteria but also from the exosymbiont - the secondary endosymbiotic host - presumably helped diatoms to adapt to ecologically new niches. For example, diatoms display a ornithine-urea cycle, which was thought to be restricted to animals, that results from the tight coupling between exosymbiont and bacterial derived proteins, and plays an important role in the metabolic response of diatoms to episodic nitrogen availability (Allen et al., 2011). Comparative genomics of the first two sequenced diatoms, *Thalassiosira pseudonana* (Armbrust et al., 2004) and *Phaeodactylum tricornutum* (Bowler et al., 2008), representative of the centric and pennate groups, revealed over 300 bacterial genes in both diatoms, some hinting to their endosymbiotic origin but also posterior acquisition through horizontal gene transfer, constituting a major driving force of diatom evolution.

Divergence. Insights from microfossils suggest that the emergence of diatoms took place in the Triassic period, 252 to 201 Myr ago (Sims et al., 2006) although the earliest well-preserved diatom fossils come from the Early Jurassic, around 190 Myr ago. During the Cretaceous period, between 145 and 65 Myr ago, diatoms are believed to have played an important role in the carbon cycle along with other photosynthetic organisms such as dinoflagellates and coccolithophorids. This was accompanied by an increase of oxygen in the surface waters, and decrease in iron availability that is favorable to algal growth. Multipolar centrics, a major lineage of Bacillariophyta, diverged in the early Cretaceous (**Figure 1.15**). At the end of the Cretaceous, about 65 Myr ago, the mass extinction induced a loss of 85% of all species of Earth including that of marine phytoplankton. However, diatoms managed to survive and began to colonize offshore areas, like the open ocean (Armbrust, 2009). This period saw the emergence of araphid pennates as indicated in fossil records (Sims et al., 2006). The emergence of raphid pennates occurred 30 Myr ago, equipped with the ability to glide along surfaces hence expanding the ecological niches greatly. New estimates of diatom diversity in the Cenozoic, between 65 Myr ago and present, report major increases near the Eocene/Oligocene boundary (30 Myr ago) as well as mid Miocene (15 Myr ago), and associate warmer oceans with lower diatom diversity (Lazarus et al., 2014).

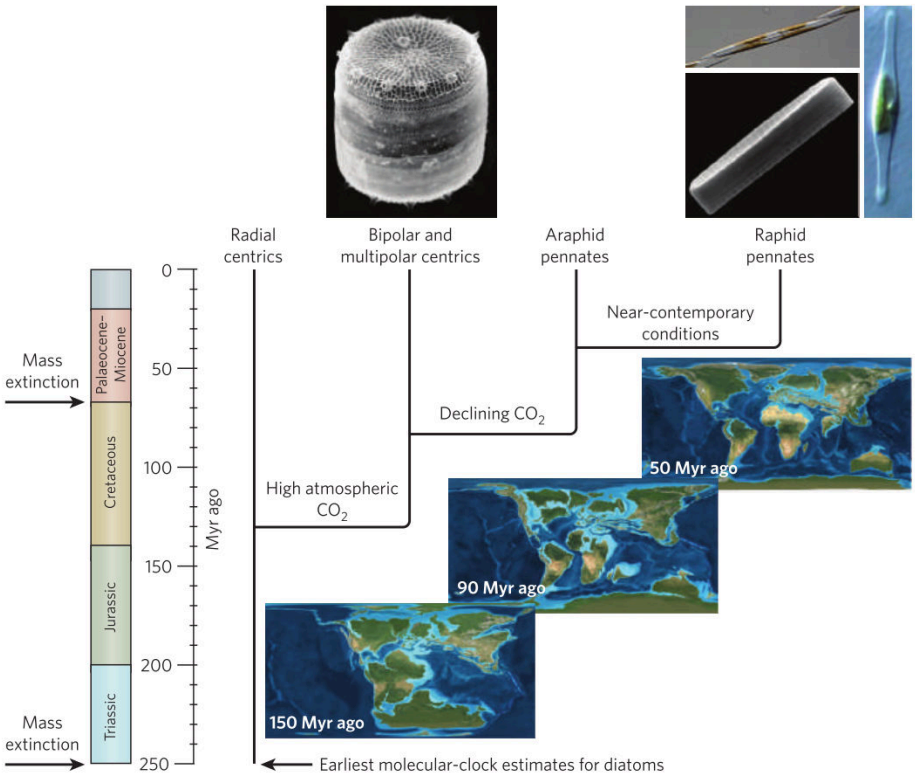


Figure 1.15. Estimated timing of divergence of the four major diatom lineages and coincident events in Earth's history.
Shown above two of the branches are images of the four species for which the whole genome sequence is available: the multipolar centric *Thalassiosira pseudonana*, and the raphid pennates (from left to right) *Pseudo-nitzschia multiseries*, *Fragilariopsis cylindrus* and *Phaeodactylum tricornutum*. Maps are palaeographic reconstructions of continent locations during the emergence of the diatom lineages. Shallower depths in the ocean are indicated by lighter blues (Armbrust, 2009).

Molecular phylogeny of diatoms. The evolution of diatoms can also be investigated thanks to molecular approaches (see Annex A). The diatom phylogenic tree consensus inferred from the 18S rDNA-gene region by Kooistra (**Figure 1.16**) reveals the following patterns: *Ellerbeckia sol* and its relatives form the basal radial centric clade, sister to the radial centrics. Basal ramifications of radial centrics are badly resolved, however they divide in a few well-supported radial centric lineages, containing chain-forming diatoms or solitary planktonic cells. The highly diverse clade with multipolar centrics and pennates emerges from the basal centric ramification. Well-supported lineages emerge in the paraphyletic multipolar centrics (group that does not include all descendants of a single common ancestor), displaying the ability to produce mucilage from the valve apical pore. All pennates form a well-supported clade, divided in two lineages. First, the araphid pennates, characterized by elongated valves and the ability to form colonies. Second, the raphid

pennates that possess a raphe slit and are the only monophyletic group. Overall, it seems that the raphids evolved from araphids, and that the araphids are derived from the centrics.

This classification does come with some controversy with respect to other morphological and molecular phylogenies, which will not be discussed here for the sake of length. This tree was detailed as it is based on the 18S marker used in this thesis, contains most of the species named in this manuscript, and the nomenclature of Radial Centric, Polar Centric, Araphid Pennate and Raphid Pennate was adopted to assign the diatom molecular data.

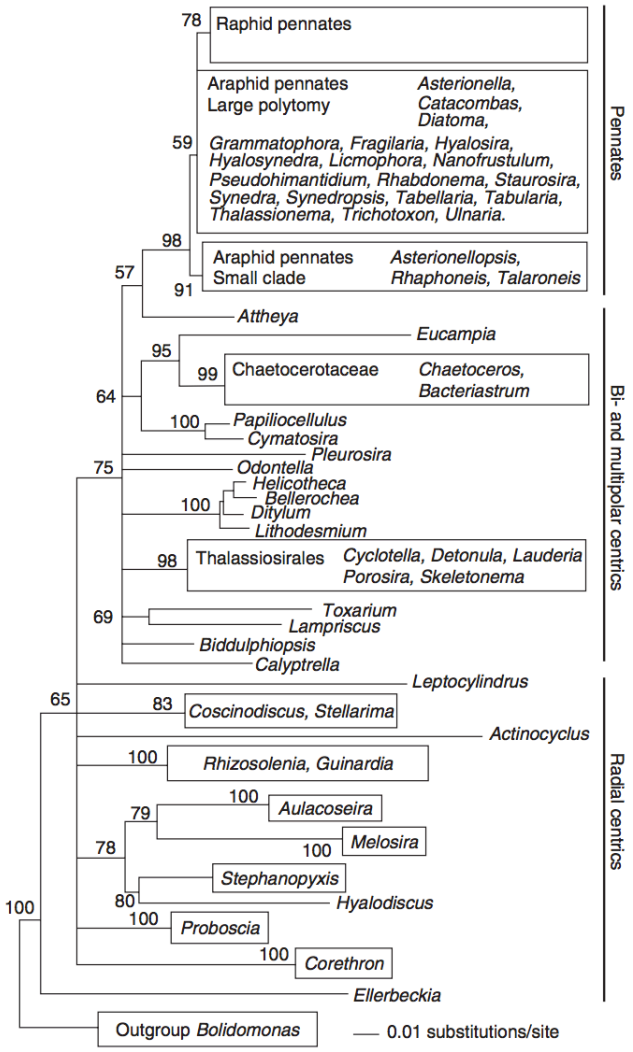


Figure 1.16. Diatom Neighbor joining phylogeny.
Inferred from maximum likelihood pair-wise distances among nuclear SSU rDNA sequences of various diatom species Sequences of Bolidomonas spp. were included as outgroup. Maximum likelihood calculations were constrained with substitution rate parameters: ADC = 0.9, ADG = 2.4,ADT = 1.2, CDG = 1.1, CDT = 3.8 versus GDT set to 1.0. Assumed proportion of invariable sites along the alignment = 0.4 and a gamma distribution of rates at variable sites with shape parameter (alpha) = 0.5. Assumed nucleotide frequencies: A = 0.25, C = 0.18, G = 0.25, T = 0.32. Bootstrap values (1000 replicates) have been generated using the same settings; values associated to minor clades to the right have been omitted (Kooistra et al., 2007).

1.2.1.5. Diatom biogeography and assemblages

Biogeography. Diatoms are prolific phototrophic organisms, and inhabit the open ocean, polar waters, tropical waters, all fresh water areas, soil, snow, and even glacial ice, for which they have developed different adaptations to survive. These adaptations include strong silica frustules for grazing protection, energy storage under the form of chrysolaminarin and oil, as well as physiological adaptations to maintain positive buoyancy, such as vacuole content less dense than seawater, achieved by containing light ions. But why do diatoms live where they do? Diatoms are drifting organisms, and we poorly understand to what extent their spatial distribution is controlled by local environmental selection, or if it is limited by dispersal. Cermeño (Cermeño et al., 2009) analysed fossil diatom assemblages over 1.5 million years from the world's ocean, to show that diatoms are not only limited by dispersal. What this paper suggests is that environmental selection - rather than dispersal - dominates diatom community structure and biogeography, contrary to macroscopic organisms for which geographic isolation is a factor of speciation. Diatom assemblages are additionally determined by the intricate synergy between bottom-up factors, such as light and nutrient supply, for which we have a fair understanding (Falkowski, 1998), and top-down factors such as grazing, for which our knowledge remains scarce (Smetacek, 2011).

Diatoms dominate phytoplankton communities in well-mixed coastal and upwelling regions, as well at the sea ice edge in polar regions, as long as sufficient light, inorganic nitrogen, phosphorus, silicon and trace elements are available to sustain their growth (Morel, 2003). In particular, diatoms can be at the source of massive algal proliferations called "blooms" that last a few weeks, that are often triggered by bottom up factors such as incident irradiance, nutrient availability and surface mixed layer shallowing (Platt et al., 2009). Blooms typically occur in the early spring and last until late spring or early summer. This seasonal event is characteristic of the temperate North Atlantic Ocean, sub-polar, and coastal waters. Yet, blooms cannot be explained just by the fact that diatoms have a superior environmental tolerance or more efficient nutrient uptake systems relative to other photosynthetic organisms. If this were the case, why would diatom species dominating blooms experience less grazing mortality than do co-occurring diatom species

(Assmy et al., 2007; Strom et al., 2007)? Several hypotheses involve top-down effects from biotic interactions between diatoms and other members of the plankton.

1.2.2. The diatom social network

Diatoms are an extremely good case study for biotic interactions. They are pivotal in marine microbial communities and are known to interact with numerous other organisms in the plankton. These interactions can provide insights about why diatoms can thrive in oligotrophic waters, how they can outcompete other organisms in eutrophic conditions and ultimately how these interactions shape plankton communities. For a definition of "biotic interactions", please refer to Annex B.

1.2.2.1. A large variety of interactions

- Symbiosis

We restrict the meaning of symbiosis to close mutualistic relationships, whereby two species that interact both individually benefit from the association (Paracer et al., 2000).

Diazotrophs. A highly mutually beneficial interaction involving diatoms is that known to occur with diazotrophic prokaryotes, such as the heterocystous cyanobacteria *Richelia intracellularis* and *Calothrix rhizosoleniae,* observed in low nutrient oligotrophic oceans. *Richelia*, along with *Trichodesmium*, is believed to be a major prokaryotic fixer of dinitrogen gas (N_2) in the world's tropical and subtropical oceans (Carpenter et al., 2002). *Richelia intracellularis* converts dinitrogen gas in ammonium and then supplies the diatom with fixed bioavailable nitrogen compounds essential for metabolism (Rai et al., 2002; Foster et al., 2011). In these cases the diatom serves as a protective host as the cyanobacteria lives inside the diatom. Another less studied symbiosis engages the chain-forming pennate diatom, *Climacodium frauenfeldianum* and a unicellular cyanobacterium similar in morphology to the free-living diazotroph, *Crocosphaera watsonii* (Foster et al., 2011). These interactions are referred to as "DDA": Diatom-Diazotroph Associations (**Figure 1.17**). *Richelia* lives as an endosymbiont between the cell wall and the frustule of diatoms such as *Hemiaulus*, *Rhizosolenia* and *Bacteriastrum*, whilst *Calothrix* lives externally attached to *Chaetoceros* spp. (Fogg, 1982; Villareal, 1994) and successive efforts to molecularly identify the partners,

using *nifH*, 16S rRNA and *hetR* sequences, has revealed the phylogenetic relationships between different diazotrophs (Janson et al., 1999; Foster et al., 2006). Comparative genomics studies of two obligate and facultative symbiont strains show that the location of the symbiont (intracellular or extracellular), and its dependency on the host are linked to the evolution of the symbiont genome, especially in nitrogen metabolism, assimilation genes, and genome reduction (Hilton et al., 2013).

Diatoms from the Rhopalodiacean family also contain an endosymbiont of cyanobacterial origin, named the "Spheroid body" that is obligate. Diatoms such as *Rhopalodia* and *Epithemia* can grow in nitrogen-poor habitats, suggesting that the endosymbiont fixes atmospheric nitrogen. The recent sequencing of the spheroid body genome was found to be considerably reduced compared with its close free living relatives and depleted in key metabolic capacities such as photosynthesis, making it completely dependent on its host (Nakayama et al., 2014). Such intricate associations are not always mutualistic; instead, some are highly detrimental to the diatom host, in which case the interaction is then referred to as parasitism.

- Parasitism/Pathogenesis

Parasitism is described as a common consumer strategy, whereby parasites generally feed on only one prey, are smaller than their host and do not usually kill the host, unlike parasitoids (Lafferty et al., 2002). Parasitic epidemics frequently follow diatom blooms in lakes worldwide, sometimes affecting over 90% of the population.

Zoosporic parasites. In the marine ecosystem, the ecological role of parasites infecting diatoms is poorly understood. Knowledge about marine diatom zoosporic pathogens is summarized in (Scholz et al., 2016), suggesting marine diatom diseases may have significant impacts on the ecology of individual diatom hosts, but also at the level of the community. Zoosporic parasites are facultative or obligate and produce spores as they infect the host. Known diatom parasites involve chytrids, aphelids (*Pseudaphelidium drebesii* parasite of *Thalassiosira punctigera*), stramenopiles - including oomycetes, labyrinthuloids, and hyphochytrids - (*Ectrogella perforans* parasite of *Licmophora hyalina*), parasitic

dinoflagellates (*Paulsenella vonstoschii* parasite of *Streptotheca tamesis* diatom), cercozoans (*Cryothecomonas aestivalis* parasite of *Guinardia delicatula*) and phytomyxids (*Phagomyxa bellerocheae* parasite of *Bellerochea malleus*). The review concludes that diatom zoosporic parasites are much more abundant in the marine ecosystem than what the available literature reports (**Figure 1.17**).

Gsell reported an interesting case of diatom - parasitic interaction in 2013 (Gsell et al., 2013). The study investigated the susceptibility to infection of seven different genotypes of the spring bloom freshwater diatom *Asterionella formosa* by a single genotype of the chytrid parasite *Zyghorhizidium planktonicum* across five environmentally relevant temperatures. The results suggested that the thermal tolerance range of the parasite genotype was narrower than that of its host, providing the diatom with a "cold" and "hot" thermal refuge in which it was not infected by the parasite. The reaction to parasitism was host-genotype specific and varied with temperature so much so that no host genotype would out-compete the others across all temperature ranges. The authors inferred that thermal variation plays a role in the maintenance of diatom diversity in disease-related traits. This also highlights the importance of environmental factors in the establishment - or not - of an interaction. Host parasite specificity and environmental factors such as temperature can impact diatom diversity, survival and, consequently, their role in community structure. Other environmental parameters such as nutrient availability can trigger diatom interactions whereby organisms compete for similar resources.

- Competition

The diversity of planktonic organisms in a given environment has puzzled scientists for a long time, raising the question of how so many different plankton species could stably coexist in a given environment, especially when they are occupying the same niche and in need of the same resource, a mystery also known as the "paradox of the plankton" (Hutchinson, 1959). Some - like Hardin - state that species do not cohabit but rather adhere to the "Competitive exclusion principle" according to which two species competing for the same resource cannot stably live together, as long as other ecological factors remain constant (Hardin, 1960).

Intra – taxa competition. Diatoms compete for nutrient resources, for instance with other diatoms. A recent metatranscriptomic study performed on the East Coast of the USA revealed that similar marine diatom species, *Skeletonema* spp. and *Thalassiosira rotula,* utilize resources differently thereby enabling their coexistence in the same parcel of water, despite similar requirements in nitrogen and phosphorus. The former favoured uptake of inorganic nitrogen sources (nitrate and nitrite), whilst the latter favoured the utilization of nitrogen from organic sources, such as amino acids (Alexander et al., 2015). Competition amongst diatoms can also result from the coupling of nutrient limitation, such as silica-limited environments, and physical factors such as temperature. Different diatom species grow unequally with respect to those co-varying factors, suggesting a specific niche adaptation, as was also shown in freshwater diatoms (Shatwell et al., 2013).

Inter – taxa competition. Diatoms also compete with radiolarians, another silica biomineralizing group of plankton. The diatom expansion 65 Myr ago has been attributed to their superior competitive ability for silicic acid uptake relative to radiolarians, the latter experiencing a reduction in weight of their tests (Harper et al., 1975). However, as the size reduction of radiolarian tests was insufficient to explain diatom expansion, strong long term erosion of continental silicates has been proposed as a significant co-factor of diatom growth (Cermeño et al., 2015). Diatoms also compete for resources with **dinoflagellates** through chemically-mediated processes detailed in the Allelopathy section below. Whereas diatoms compete for resources to grow and survive, other organisms at higher trophic levels need diatoms to grow by feeding upon them through predation.

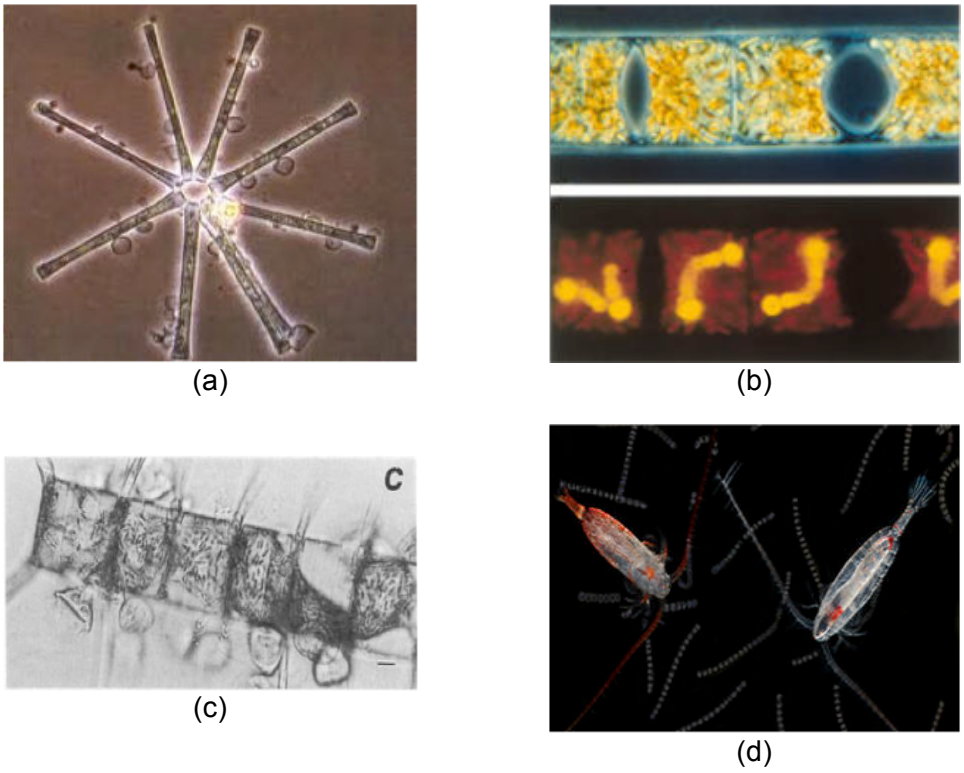


Figure 1.17. Diatoms are involved in a large variety of interactions.
(a)Parasitism by the Chytrid (*Zygorhizidium planktonicum*) on *Asterionella formosa* (NIOO) (b) Symbiosis between the diatom *Hemiaulus sp.* (red: chloroplast) and the N2 fixing cyanobacteria, *Richelia* (Yellow) (Foster et al., 2011). (c) *Chaetoceros costatum* and epiphytic ciliate *Vorticella oceanica* (Sachiko et al., 1996) (d) Copepods feeding on *Skeletonema* diatom (Hicks et Coull, 1983).

- Predation

Diatoms are often referred to as the “pastures of the sea” (Smetacek, 2001). Indeed, out of the myriad of mechanisms that can induce phytoplankton mortality or remove phytoplankton biomass, such as viral lysis or sinking, predation is quantitatively dominant (Calbet et Landry, 2004) maintaining ratios of primary producers to herbivores very low and is therefore a structuring factor in the plankton (Sherr and Sherr, 2009). Unlike parasites that also feed on diatoms, it is generally assumed that predators feed on several species (not one), tend to be bigger than the prey, and tend to kill it (Lafferty et al., 2002).

Metazoan predators such as copepods (crustaceans) presumably exercise strong pressure on diatoms by feeding on them (Lebour, 1922; Marshall, 1955; Campbell et al., 2009). The

classic pelagic food web involves a trophic linkage between diatom blooms, copepod production, and fish (Runge, 1988; Legendre 1990). Numerous feeding experiments have investigated the coevolution between copepods and diatoms. Some adaptations are mechanical: copepods modify their feeding tools (Itoh, 1970; Michels et al., 2012), in response to which diatoms adjust their protecting frustules, leading to an arms race that fuels evolutionary processes (Hamm et al., 2007). Some diatoms that dominate blooms experience less grazing mortality than do co-occurring species (Assmy et al., 2007; Strom et al., 2007): it was shown that in the presence of preconditioned media that contained herbivores, diatoms develop grazing resistant morphologies such as increased cell wall silicification (Hamm et al., 2003; Pondaven et al., 2007). Hence, the cell wall provides not only a "constitutive mechanical protection" for the cell but also a plastic trait that responds to grazing pressure.

Other adaptations are physiological: the existence of a mismatch between temperature optima for growth of diatoms relative to growth of potential predator is a strategy to escape predation pressure (Rose et al., 2007). Finally, the production of defensive chemicals and allelopathic molecules targeted towards predators is thought to contribute to diatom success, although largely debated, as detailed below. The classic food web view was challenged in the early 1990's (Kleppel et al., 1991). It was suggested that copepods rather feed preferentially on microplankton such as ciliates and dinoflagellates, supported by evidence that diatoms were nutritionally insufficient for copepod growth (Jónasdóttir et al., 1998). Additional arguments favour low copepod grazing pressure during blooms: the copepods inability to track diatoms overwinter and the existence of grazing from heterotrophic dinoflagellates.

Heterotrophic dinoflagellates are unicellular non-pigmented phagotrophic microplankton measuring between 20 and 100 microns, and are probably the highest consumers of bloom-forming diatoms, more than copepods and other mesozooplankton (Sherr and Sherr, 2007). They can comprise more than 50% of microzooplankton biomass in diatom blooms, represented by thecate dinoflagellates (armored, like *Protoperidinium* spp.) and athecate (*Gymnodinium* spp.). They exert a constant predation pressure on diatoms, due to asexual

reproduction - therefore a rapid matching when resource abundance increases - but also by their capacity to grow on diverse prey, therefore surviving in non-bloom conditions to better proliferate when diatoms bloom (Strom, 2008). Attempts to compare the dinoflagellate and copepod pressures on diatom communities have been done in South Korean coastal waters. Dinoflagellates (*Protoperidinium bipes*) consumed 0.1-3.4%*H-1 of diatom biomass, whereas copepods (*Acartia spp.*) removed less than 0.2%*H-1 diatom biomass, rather focusing on herbivore ingestion and relieving diatoms from grazing pressure (Jeong et al., 2004). Experimental simulation of trophic interactions among omnivorous copepods, heterotrophic dinoflagellates and diatoms also suggest that dinoflagellates play a central role in the lower trophic levels of marine food webs by consuming diatoms and then serving as a quality food source for copepods (Chen et al., 2011).

- Interactions with unclear benefits

Biotic interactions cannot always be classified easily. In the following section, I focus on interactions, either beneficial or detrimental, involving micro-organisms, mediated by a form of physical attachment/proximity and for which the consequences for the organisms are unclear.

Diatom Host-Associated Microbes. With the arrival of next generation sequencing, host-associated microbiomes have attracted much attention because microbial diversity becomes easier to characterize. By producing a thick organic rich biofilm, several diatoms are covered with bacteria, and it was estimated that 20% of bacterioplankton may actually be associated to algae (Azam et al., 1998). The *Pseudo-nitzschia* diatom genus is well studied because certain representatives produce domoic acid, a toxin causing disease in shellfish and ultimately in humans following ingestion, and produced in the presence of zooplankton (Tammilehto et al., 2015). In particular, it was shown that toxin-producing *Pseudo-nitzschia* had a lower associated microbial diversity than non-toxic species, suggesting that some members of the microbiota were truly specific to their native host, as they became parasitic when administrated to a foreign host (Sison-Mangus et al., 2014). Earlier, Kaczmarska et al. (2005) used scanning electron microscopy to show that the bacterial community present on *Pseudo-nitzschia multiseries* in culture was very different from the one present on diatoms

in native seawater, highlighting the dynamic aspect of the microbial community. Amin et al. (2012) authored an excellent review on interactions between diatoms and bacteria, and suggested that the prominent bacterial communities associated with diatom cultures are Proteobacteria, Bacteroides (the main heterotrophic phyla associated with diatoms), and in particular Sulfitobacter, Roseobacter, Alteromonas and Flavobacterium. The respective benefit for each organism is not always straightforward though the small size of bacteria is today an obstacle that can be overcome to improve our knowledge about bacteria/phytoplankton interactions.

The first step is often to try and characterize the interactions based on external features, however understanding the mechanism of interaction is necessary to uncover the true nature of the association. In some cases diatom interactions mediated by chemical components have been better described and have enabled the understanding of many opaque associations.

- Chemical bouquets of interactions

Allelopathy at large is a biochemically-mediated interaction in which one organism can influence growth, survival, and reproduction of another organism. The effects can be either beneficial (positive allelopathy) or detrimental (negative allelopathy). These chemical signals can influence species interactions in the plankton, which is well illustrated in phytoplankton (Legrand et al., 2003), and particularly in diatoms. I will not discuss here the distinction between allelopathy and competition (Willis, 2007), but describe below a range of biotic interactions that have been identified as being chemically mediated, including competition for resources.

Allelopathy in response to copepod grazing. Copepods graze on diatoms, and there has been much debate about whether or not diatoms are a good food source for copepods, in what is known as the “Diatom-Copepod Paradox” (Harvey et al., 1935). In the early 1990’s it was discovered that diatom-derived compounds (simple aldehydes) could decrease copepod egg hatching success from the usual 90% to 12% (Ban et al., 1997; Miralto et al., 1999), challenging the classical view of marine food webs wherein energy flows from diatoms to fish by means of copepods (along with the discovery of high grazing rates by

dinoflagellates). Further studies discovered a myriad of polyunsaturated aldehydes named "PUAs", in the diatoms *Thalassiosira rotula* and *Skeletonema costatum,* which are released within seconds after mechanical crushing of the diatoms, up to 5 fmol of PUA per cell within two minutes (Pohnert, 2000).

PUAs are the breakdown products of the oxidative transformation of membrane fatty acids. They have been found to inhibit cell proliferation and cell division, and to induce phagocytosis and apoptosis. While the teratogenic effect of PUAs does not appear to deter the herbivores from feeding, it does impair recruitment when female copepods are fed with PUA-rich diets, compromising the development of the offspring (Ianora et al., 2004). Similar effects were illustrated with diatom PUAs on the polychaetes *Arenicola marina*, and *Nereis virens*. Recent transcriptome analysis using expressed sequence tags from the copepod *Calanus helgolandicus* showed that two days of feeding with the PUA-producing diatom *Skeletonema marinoi* activated a cellular stress response in the grazer, and impaired the reproductive and developmental processes in copepods such as gametogenesis, embryogenesis, egg viability and sex differentiation (Carotenuto et al., 2014). Investigation of PUAs in a diatom bloom in the North Adriatic Sea (Ribalet et al., 2014) found that PUA concentrations correlated with diatom cell lysis, however PUA producing diatoms were not reported in cases of high PUA concentrations (2,53nM). This thickens the plot, and raises new questions: is there another PUA-producing organism, or do the enzymes responsible for PUA generation retain their activity extracellularly after diatom cell lysis? On the contrary, in 2002 Irigoien presented *in situ* data from which he could not observe any negative relationship between copepod egg hatching success and diatom biomass, or dominance, thus defending the classical view of food web dynamics. This was reinforced by Flynn (Flynn et al., 2009) and Irigoien as co-author, who evaluated the effect of diatoms on copepod reproduction, and found that the aldehyde-induced negative effect could not be considered as a diatom defence mechanism against copepods, including the inhibition of egg viability.

Allelopathy by dinoflagellates for resource competition. The study of nearshore blooms of the dinoflagellate *Karenia brevis* proposed that allelopathic (unstable) compounds were produced to inhibit growth of phytoplankton competitors, among which diatoms (Prince et

al., 2008; Poulson et al., 2010). However, natural offshore diatom-dominated assemblages in the Gulf of Mexico seemed resistant (*Asterionellopsis glacialis*, *Skeletonema* spp.), even displaying slight stimulation of growth, results that are again more variable when brought back to the lab. The accumulation of allelopathic compounds in the water column may create an inhospitable environment for growth among competitors, although diatom response are clearly species specific. In the lab, *Karenia brevis* caused suppression of growth of *Thalassiosira pseudonana* and *Asterionellopsis glacialis* and the impact of the dinoflagellate on the competitors' physiology was reflected in the metabolomes and proteomes of both diatoms. Cellular protection responses such as altered cell membrane components, inhibited osmoregulation and increased oxidative stress were triggered (Poulson-Ellestad et al., 2014).

Allelopathy by prokaryotes. *Kordia algicida* belongs to the Flavobacteriaceae family, and its cell free filtrate was observed to display a high protease activity that induces an algicidal action in several marine diatoms such as *Skeletonema costatum, Thalassiosira weissflogii,* and *Phaeodactylum tricornutum* (Paul et al., 2011). Lee (Lee et al., 2000) demonstrated that *Pseudoalteromonas* sp. produce a high molecular weight extracellular protease that is able to inhibit the growth of the diatom *Skeletonema costatum*.

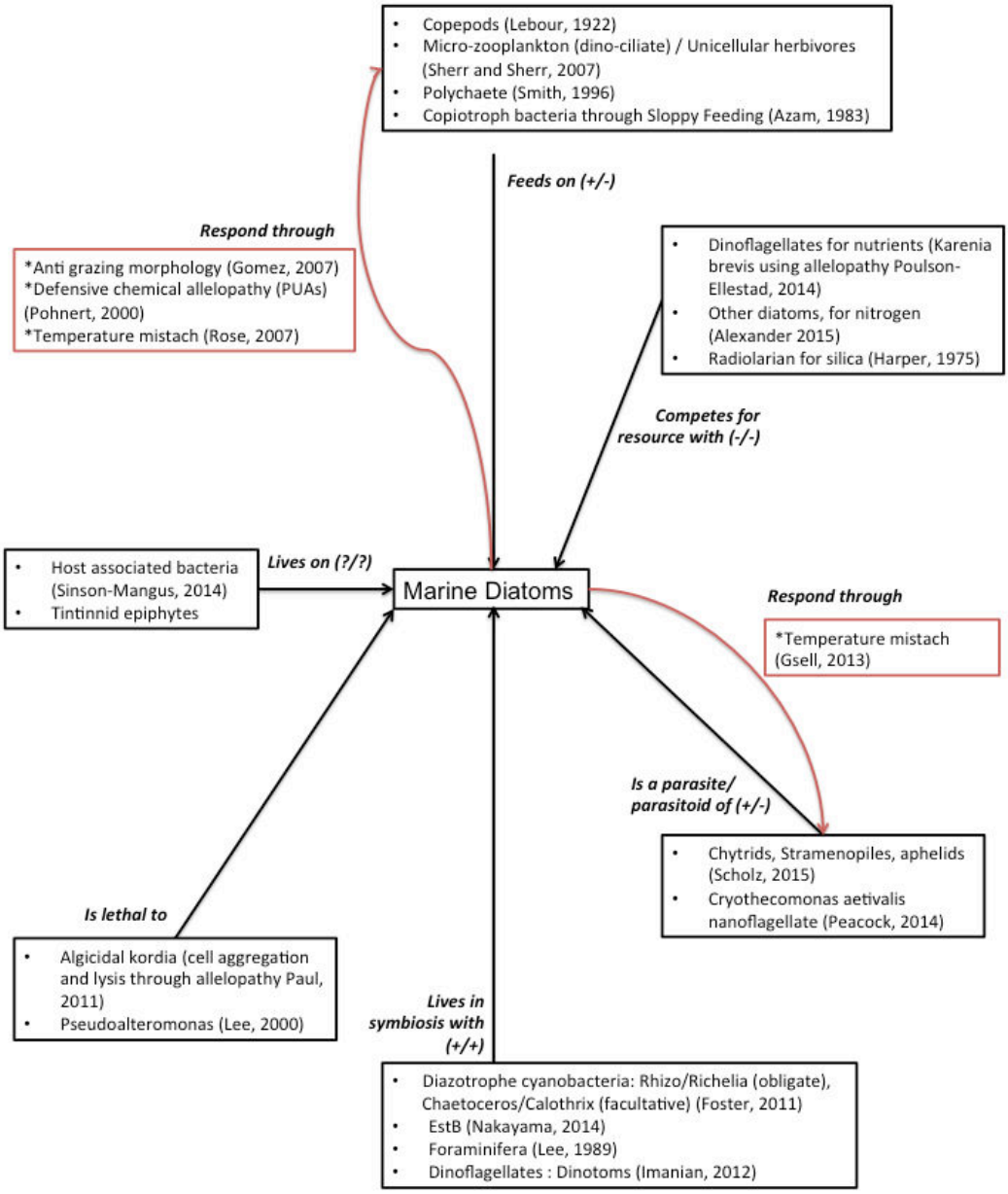


Figure 1.18. Diatoms, pivotal in marine microbial interactions.
(+/+), (+/-), (-/-) and (+/-) signs refer to the relative positive, negative, or neutral effect both interacting species can have on each other. The first term refers to the effect of the interaction of the partner of the diatom, and the second to the effect of the interaction on the diatom. See Annex B for further details.

The relative contribution of these different interactions in shaping the evolution of diatoms is essentially unstudied. In 2001, Smetacek argued that the evolution of plankton was likely ruled by protection against grazing, and not by competition for resources, and therefore

that the interpretation of blooms as being the outcome of superior environmental tolerance and resource competition among photosynthetic protists was incomplete. For him, the many different morphologies and life histories of diatoms reflected responses to specific top-down pressures such as predation, and thus that we should not adapt our understanding of the evolution of form and function in terrestrial vegetation - driven by competition for resources and resource space (bottom-up) - to phytoplankton. The complexity of an integrated view of diatom biotic interactions does not stop with the diversity of partners involved, or the different mechanisms developed. Additionally, these interactions cannot be considered as snapshots, but rather as dynamic processes, both spatially and temporally across multiple scales.

1.2.2.2. Interactions that vary through space and time

Temporal scales of diatom interactions. Diatom produce PUAs (Polyunsaturated fatty acids) in the **seconds** following the crushing of the diatom frustule induced by predation by larger grazers. In the following **hours**, the copepods will continue eating in this environment garnished with teratogenic compounds. The interaction, on the long term, will have an impact on the offspring so much so that over a few **years**, grazers will evolve to avoid eating PUA-producing diatoms. Teeling et al. (2012) investigated the bacterioplankton response to a diatom bloom in the North Sea and managed to uncover the dynamic succession of bacterial populations at the genus level. Over a few **days**, bacteria known to decompose algal-derived organic matter, such as Bacteroidetes, Gammaproteobacteria, and Alphaproteobacteria, formed distinct, successive populations controlled by algal substrate availability. Over **decades**, biotic interactions leave their imprint in the seasonal succession of plankton, an annually repeated process of community assembly that is the result of community interactions such as competition, predation, and parasitism in conjunction with abiotic control mechanisms that set the start and end of the growing season. The studies of these dynamics, by sampling regularly at a given location, or by following a prevailing current, also known as a time series study, enables scientists to examine how different organisms change in relation to one another and in relation to environmental conditions (Fuhrman et al., 2015). Over **millennia**, past endosymbiontic events and other gene transfers remain traceable in the genetic information of diatoms or their host.

Spatial scales of diatom interactions. Patches of homogeneous communities, and thereby the interactions that happen amongst them, can be observed from scales of **km** to thousands of km at a given depth and over horizontal directions. In the vertical direction, there can be significant changes over **meters**, or tens of meters or even millimeters within ephemeral microlayers (Fuhrman, 2009)**.** The physical contact between a copiotroph bacteria and the mucus of the diatom, the bacterial diazotroph encapsulated in its host, or what happens at the cell surface in general through defense and protection against agents of mortality happens over a few **micrometers**. Diatom interactions enter scales of **centimeters** when copepods feed on them. Symbiosis with cyanobacteria can form blooms measured in **kilometers**, as was reported in the subtropical North Atlantic (Carpenter et al., 1999), estimating that the N supply by N_2 fixation by the symbioses exceeded that of nitrate flux from below the euphotic zone, thus playing a significant role in the biogeochemistry of the surface ocean. Similarly, it was shown that DDAs drive a significant biological CO2 pump in tropical oceans off the Amazon River plume (Yeung et al., 2012) illustrating how biotic interactions can scale up to influence ecosystem wide phenomenon.

We therefore see that diatom interactions are diverse (**Figure 1.18**), spanning across multiple temporal and spatial scales, involving both macro- and micro-organisms, prokaryotes and eukaryotes and even viruses (that were not mentioned here). Many of these studies rely on manipulative experiments, such as co-culturing (two organisms in same medium) and cross culturing (cell free filtrate of one organisms added to the medium of the target) of potential competitors, feeding experiments to test specificity of prey and predators, and they have more recently also embraced the omics era. Transcriptomic data has been used to evaluate copepod responses to harmful diatoms, DNA barcoding has been used to analyse predator gut content (Kress et al., 2015), metabolomics has helped understand allelopathy, and genomics has helped interpret the evolution of host-symbiont gene transfers and evolution. But microbial communities are complex, and most studies provide a reductionist view, studying one, two, or in best of cases three organisms in isolation.

The need to develop holistic approaches emerged a few years ago in marine microbiology (Fuhrman, 2009), and the possibility to study organisms in their natural habitats has opened the door to novel ways of looking at community structure in the microbial aquatic world. With the broad empirical knowledge acquired on diatom biotic interactions, the remaining question is to understand **how they structure the plankton community at large spatial scales**. A dataset that would give the opportunity to go from single cell to the global ocean to characterize biotic interactions in their natural environment could be transformational for the field.

1.3. The *Tara* Oceans expedition

Over many centuries, global expeditions have led to major scientific breakthroughs, notably with the early voyages of the H.M.S. Beagle (1831–1836) and the H.M.S. Challenger (1872–1876). Ocean exploration now provides promising first steps towards understanding the role of the ocean in global biogeochemical cycles and the impact of global climate change on ocean processes and marine biodiversity. Recently, the Sorcerer II expeditions (2003–2010) (Rusch et al., 2007) and the Malaspina expedition (2010–2011) carried out global surveys of prokaryotic metagenomes from the ocean's surface and bathypelagic layer (>1,000 m), respectively. The *Tara* Oceans expedition (2009–2013) complemented these surveys by collecting a wide variety of planktonic organisms (from viruses to fish larvae) from the ocean's surface (0–200 m) and mesopelagic zone (200–1,000 m) at a global scale. Overall, *Tara* Oceans surveyed 210 ecosystems in 20 biogeographic provinces, collecting over 40,000 samples of seawater and plankton. Many questions about marine microbial organisms can be addressed in light of current technologies: what is the real nature of microbial diversity in the ocean? What are the organisms carrying out the most important functions in the ocean? What is the influence of environmental parameters? How do microbial interactions influence the marine ecology and ecosystem?

In order to address these questions, the *Tara* Oceans expedition's goal was to create a publicly available data set able to: visualize, quantify, and genetically characterize ocean biodiversity within entire plankton ecosystems, as well as to find patterns across unprecedentedly comprehensive data types.

These objectives involved sampling across multiple depths, and multiple size fractions, as well as state-of-the-art monitoring of environmental parameters (**Figure 1.19**). As a research infrastructure, the *Tara* Oceans project mobilised over 100 scientists to sample the world oceans on board a 36 m long schooner (SV *Tara*) refitted to operate state-of-the-art oceanographic equipment.

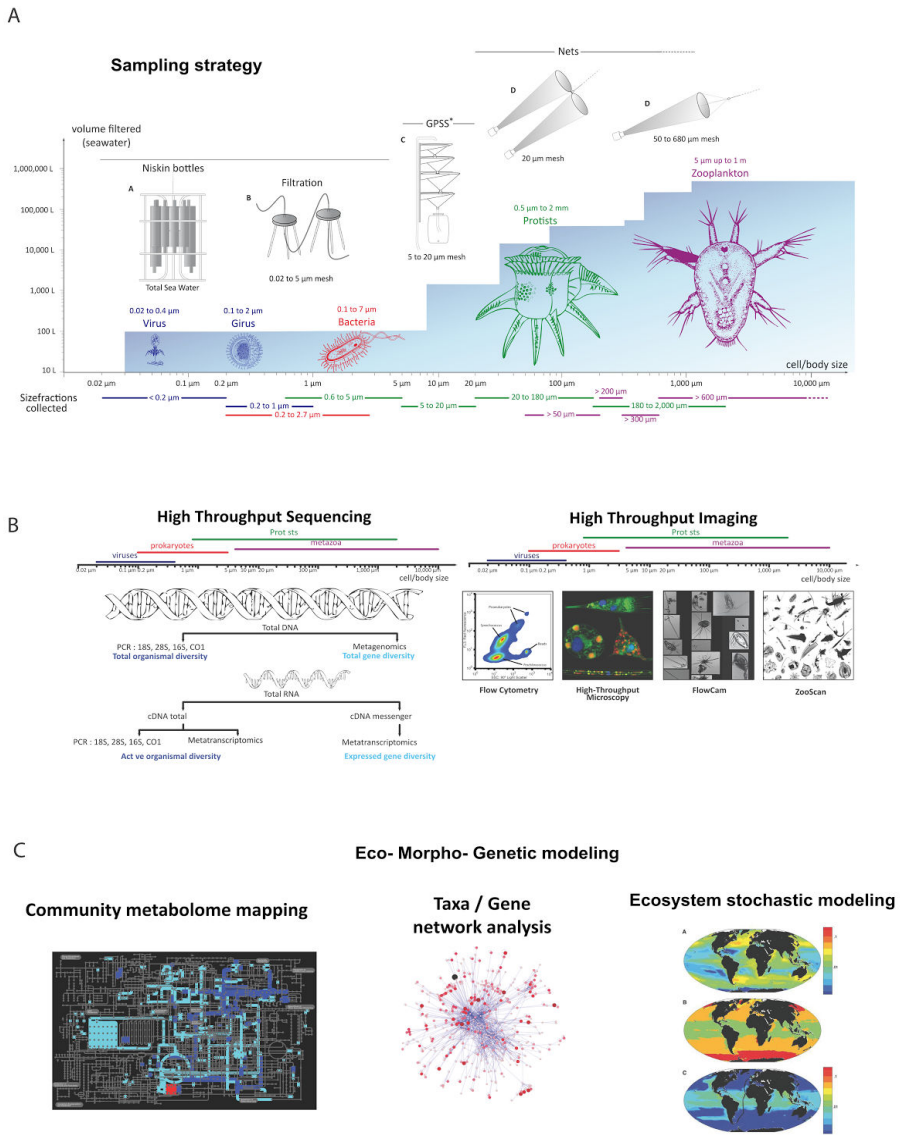


Figure 1.19. *Tara* Oceans sampling and analysis pipeline.
(A) Methods for sampling organisms by size classes and abundance. The blue background indicates the filtered volume required to obtain sufficient organism numbers for analysis. (B) Methods for analyzing samples. Data on the right are from *Tara* Oceans sampling stations. (C) 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).

The regular sampling program was designed to study a variety of marine ecosystems and to target well-defined meso- to large-scale features such as gyres, eddies, currents, frontal zones, upwellings, hotspots of biodiversity, low pH or low oxygen concentrations. A total of 210 stations were characterised at the mesoscale to provide richer environmental context for the morphological and genomic study of plankton. Each sample was processed and appropriately conserved to perform 18S rDNA metabarcoding (for an overview on metabarcoding techniques refer to Annex A), metagenomics, and metatranscriptomics sequencing (**Figure 1.20**).

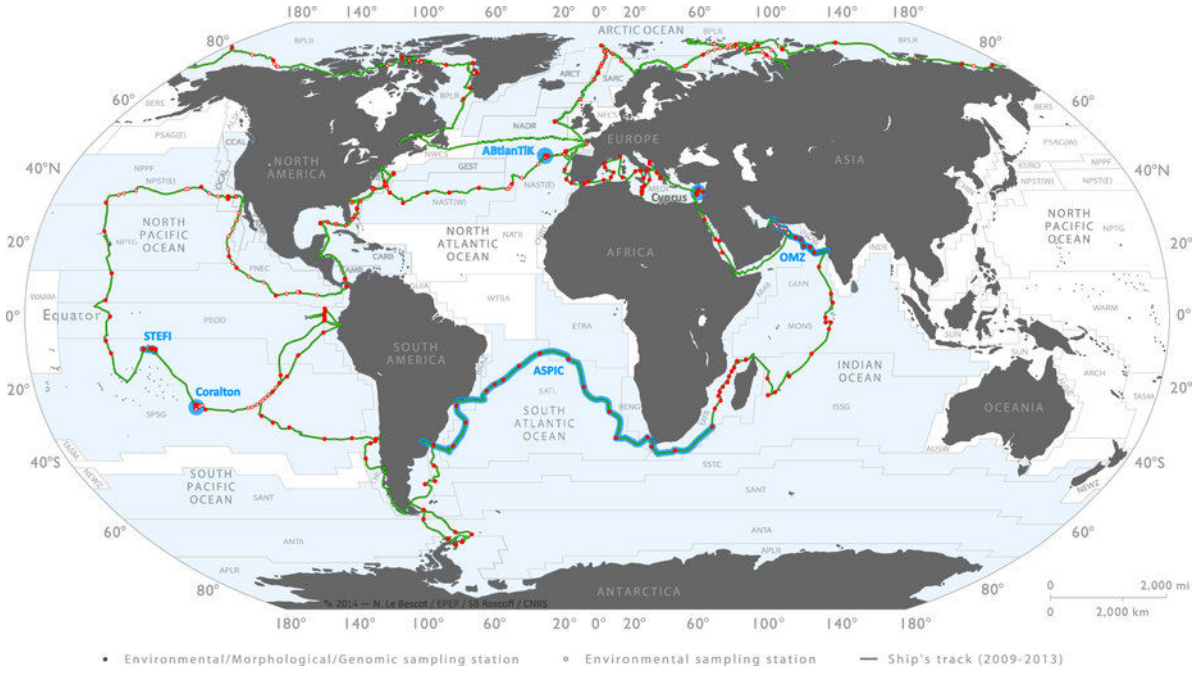


Figure 1.20. The *Tara* Oceans route from 2009 to 2013.

During the *Tara* Oceans Expeditions (2009–2013), plankton were sampled from 5–10-m thick layers in the water column, corresponding to specific environmental features that were characterized on-board from sensor measurements. Additional environmental metadata and detailed sampling procedures are available in Pesant et al., 2015 and not described here for the sake of length. Major sampling procedures relevant to this thesis are given below.

At each station, sampling was done at two depths: The surface water layer, sometimes labelled in the literature and databases as “surface”, “SRF”, “SUR”, “SURF” or “S”, was simply defined as a layer between 3 and 7 m below the sea surface. The deep chlorophyll

maximum layer, often labelled in the literature and databases as “DCM” or “D”, was determined from the chlorophyll fluorometer mounted on the Rosette Vertical Sampling System [RVSS]. The presence of a DCM may indicate a maximum in the abundance of plankton bearing chlorophyll pigments, or it may result from the higher chlorophyll content of plankton living in a darker environment. This can be assessed *a posteriori* using water samples analysed for pigments by HPLC methods and from plankton counts.

At each station, sampling was done for various size fractions. Plankton sampled during the *Tara* Oceans Expeditions cover six orders of magnitude in size and corresponds to viruses, giant viruses (giruses), prokaryotes (bacteria and archaea), unicellular eukaryotes (protists), and multicellular eukaryotes (such as metazoans). Unicellular eukaryotes, or protists, cover a broad range of cell size (0.8–2,000 µm). Nets were used for the **5–20 µm size-fraction**; plankton from the **20–180 µm size-fraction** were collected using a double plankton net with a 20 µm mesh size. Plankton from the **180–2,000 µm size-fraction** were collected using a 180 µm Bongo Net.

For each station, morphological analysis was performed for different classes of organism. This was performed through on-board and on-land FlowCams and ZooScans for quantitative recognition of organisms ranging from 20 micrometers to a few cm, light sheet and confocal microscopes for 3D imaging, on-land electron microscopes for detailed ultrastructural analyses of small protists and viruses.

The *Tara* Oceans project leverages powerful new technologies and analytical tools to develop the first planetary-scale data collection effort that links biogeography with ecology, genetics, and morphology bringing together an international community of researchers from a wide range of disciplines including marine ecologists, oceanographers, statisticians, molecular biologist, biochemists and engineers. Five foundational papers have been published in the journal Science in May 2015, and I contributed to three of them.

1.4. Thesis outline

"When we try to pick out anything by itself, we find it hitched to everything else in the universe" John Muir 1911

The goal of my thesis was to understand how diatom biotic interactions structure the planktonic community at large spatial scales, by developing new approaches based on the heterogeneous and unprecedented *Tara* Oceans dataset. This thesis is divided in four result chapters beyond this general introduction (Chapter 1), and before the conclusions (Chapter 6).

- Chapter 2: Fishing the Unknown

Chapter 2 is a **reprint** of the material as it appears in: Malviya et al., "Insights into global diatom distribution and diversity in the world's ocean." *Proceedings of the National Academy of Sciences* (2016): 201509523. On this paper I am 4th of 12 authors.

The analysis of diatom metabarcoding data (see Annex C for in-depth description of metabarcoding dataset reported in de Vargas et al., 2015, a paper on which I am 27th author) emanating from the *Tara* Oceans expedition, by Dr.Shruti Malviya, revealed that nearly 50% of the OTUs could not be assigned at the genus level. Some of them are abundant, and ubiquitous, therefore representing a strong ecological interest for community ecology. By developing a bioinformatics pipeline and experimentally validating the method taking full advantage of the global scale nature of the data set, I have successfully assigned previously "unknown ribotypes". This was complemented by manual curation of other supposed unassigned diatoms. Overall, this work allowed (1) a better taxonomic identification of diatom OTUs that represented ~ 12% of the total diatom abundance data (1,400,510 reads), (2) the extension of the ecological niche of known diatoms, and (3) a proof of concept to extend this methodology to other unassigned groups.

- Chapter 3: Global scale patterns of diatom interactions in the open ocean

Chapter 3 is composed of a detailed introduction to the methods related to co-occurrence network inference, followed by a **draft manuscript** that will be submitted to Frontiers in Microbiology **(Section 3.2)**.

Based on the large-scale co-occurrence network published in Lima-Mendez et al., 2015 (a paper on which I was 8th author; Annex D), my goal was to investigate co-occurrence patterns of diatoms and relate them to diatom biology and biogeography, with the help of graph theory. This chapter results in (1) a literature database of our current knowledge about diatom biotic interactions to detect true positives, and (2) a global scale understanding of how diatoms impact the plankton community structure by acting as repulsive segregators.

- Chapter 4: Characterisation of an abundant and widespread interaction

Chapter 4 is the long version of a **draft manuscript** that will be submitted to ISME Journal.

The wealth of the *Tara* Oceans data offers the unique opportunity to investigate microbial associations, *in situ*, at large spatial scale combining a morphological, genetic, and ecological perspective. The initial observation of a prevalent diatom biotic interaction in the South African Agulhas Current initiated the development of an integrated analysis that enabled (1) the illustration of successful data-driven experimental approaches, and (2) the characterization of a ubiquitous, specific, and fragile association between a diatom and a heterotrophic ciliate. The South African Agulhas Current and its impact on plankton dispersal is studied in Villar et al., 2015 on which I am 33rd author (Annex E).

- Chapter 5: Single cell genomics to explore diatom biotic interactions

Chapter 5 explores the use of single cell genomics to propose a holistic approach to the study of diatom biotic interactions *in situ*, extending the case study to diatom interactions with parasites and nanoflagellates.

- Annexes

Annex A : definition of biotic interactions.

Annex B: detailed explanation of the contribution of genomics to the study of microbial diversity.

Annex C : the pipeline used to process metabarcoding data used in this thesis is detailed in de Vargas et al., 2015, a paper on which I was 27th author out of 55.

Annex D : the pipeline used to compute the plankton co-occurrence network interpreted in Chapter 3 is available in Lima Mendez et al., 2015, a paper on which I was 8th author out of 51.

Annex E : the oceanic regime of the Agulhas Current mentioned in Chapter 4 is avaiable in Villar et al., 2015, a paper on which I was 33rd author out of 54.

When the text cites "**Table S** " (bold, underlined), this refers to online material provided with the manuscript.

Chapter 2: Fishing the unknown

Chapter 2 is a reprint of PNAS paper on which I am fourth author out of 12.

I was in charge of the paragraph dedicated to unassigned sequences, by performing the manual assignment of the top 100 unassigned barcodes with Lucie Bittner, thereby improving the assignment of those most abundant unassigned barcodes.

I was in charge of developing the appropriate bioinformatics pipeline and performing the experiments in order to assign cosmopolitan ribotypes that could not be assigned to a known diatom even after manual curation. The methodology is available under the "Reassignment of unknown diatom ribotypes" section of Supporting Information and will be extended to ribotypes in the whole data set that cannot be assigned, in collaboration with Sarah Romac.

I was in charge of analyzing and producing **Figure 5** on the comparison between V9 barcodes and cell abundance based on Light Microscopy counts data provided by Eleonora Scalco.

Insights into global diatom distribution and diversity in the world's ocean

Shruti Malviya[a,1], Eleonora Scalco[b], Stéphane Audic[c], Flora Vincent[a], Alaguraj Veluchamy[a,2], Julie Poulain[d], Patrick Wincker[d,e,f], Daniele Iudicone[b], Colomban de Vargas[c], Lucie Bittner[a,3], Adriana Zingone[b], and Chris Bowler[a,4]

[a]Institut de Biologie de l'École Normale Supérieure, École Normale Supérieure, Paris Sciences et Lettres Research University, CNRS UMR 8197, INSERM U1024, F-75005 Paris, France; [b]Stazione Zoologica Anton Dohrn, 80121 Naples, Italy; [c]CNRS, UMR 7144, Station Biologique de Roscoff, 29680 Roscoff, France; [d]Institut de Génomique, GENOSCOPE, Commissariat à l'Énergie Atomique et aux Énergies Alternatives, 91057 Évry, France; [e]UMR 8030, CNRS, CP5706, 91057 Évry, France; and [f]UMR 8030, Université d'Evry, CP5706, 91057 Évry, France

Edited by Paul G. Falkowski, Rutgers, The State University of New Jersey, New Brunswick, NJ, and approved January 26, 2016 (received for review May 14, 2015)

Diatoms (Bacillariophyta) constitute one of the most diverse and ecologically important groups of phytoplankton. They are considered to be particularly important in nutrient-rich coastal ecosystems and at high latitudes, but considerably less so in the oligotrophic open ocean. The *Tara* Oceans circumnavigation collected samples from a wide range of oceanic regions using a standardized sampling procedure. Here, a total of ∼12 million diatom V9-18S ribosomal DNA (rDNA) ribotypes, derived from 293 size-fractionated plankton communities collected at 46 sampling sites across the global ocean euphotic zone, have been analyzed to explore diatom global diversity and community composition. We provide a new estimate of diversity of marine planktonic diatoms at 4,748 operational taxonomic units (OTUs). Based on the total assigned ribotypes, *Chaetoceros* was the most abundant and diverse genus, followed by *Fragilariopsis*, *Thalassiosira*, and *Corethron*. We found only a few cosmopolitan ribotypes displaying an even distribution across stations and high abundance, many of which could not be assigned with confidence to any known genus. Three distinct communities from South Pacific, Mediterranean, and Southern Ocean waters were identified that share a substantial percentage of ribotypes within them. Sudden drops in diversity were observed at Cape Agulhas, which separates the Indian and Atlantic Oceans, and across the Drake Passage between the Atlantic and Southern Oceans, indicating the importance of these ocean circulation choke points in constraining diatom distribution and diversity. We also observed high diatom diversity in the open ocean, suggesting that diatoms may be more relevant in these oceanic systems than generally considered.

biodiversity | diatoms | metabarcoding | Tara Oceans | choke points

Diatoms are single-celled photosynthetic eukaryotes deemed to be of global significance in biogeochemical cycles and the functioning of aquatic food webs (1–3). They constitute a large component of aquatic biomass, particularly during conspicuous seasonal phytoplankton blooms, and have been estimated to contribute as much as 20% of the total primary production on Earth (4–6). They are widely distributed in almost all aquatic habitats, except the warmest and most hypersaline environments, and can also occur as endosymbionts in dinoflagellates and foraminifers (7).

Because of their complex evolutionary history (8), diatoms have a "mix-and-match genome" (3) that provides them with a range of potentially useful attributes, such as a rigid silicified cell wall, the presence of vacuoles for nutrient storage, fast responses to changes in ambient light, resting stage formation, proton pump-like rhodopsins, ice-binding proteins, and a urea cycle (9). In general, planktonic diatoms seem well-adapted to regimes of intermittent light and nutrient exposure; however, they are particularly common in nutrient-rich regions encompassing polar as well as upwelling and coastal areas (10), highlighting their success in occupying a wide range of ecological niches and biomes. The quantification of diatom diversity and its variations across space (and time) is thus important for understanding fundamental questions of diatom speciation and their tight coupling with the global silica and carbon cycles (8, 11), as well as for understanding marine ecosystem resilience to human perturbations.

Estimations of the numbers of diatom species vary widely, from a low of 1,800 planktonic species (12) to a high of 200,000 (13). Most recent estimates range from 12,000 to 30,000 species (14, 15). But such global estimates are confounded by the fact that most studies are focused toward understanding the patterns of diversity in a particular diatom genus at a local or regional scale (e.g., refs. 16–18). Furthermore, as evidenced from the Ocean Biogeographic Information System (OBIS) database, although diatom distributions have been explored extensively in numerous studies, they have predominantly focused on the Northern Hemisphere (19, 20).

Characterization of diatom diversity requires accurate and consistent taxon identification. Morphological analyses alone fail to provide a complete description of diatom diversity so complementary investigations are often performed to provide a uniform means of standardization (e.g., ref. 21). During the past decade, the introduction of DNA sequence analysis to systematics has facilitated the discovery of numerous previously undescribed taxa, revealing distinct species identified by subtle or no morphological variations (e.g., ref. 22). Allozyme electrophoresis (23), DNA fingerprinting

Significance

Diatoms, considered one of the most diverse and ecologically important phytoplanktonic groups, contribute around 20% of global primary productivity. They are particularly abundant in nutrient-rich coastal ecosystems and at high latitudes. Here, we have explored the dataset generated by *Tara* Oceans from a wide range of oceanic regions to characterize diatom diversity patterns on a global scale. We confirm the dominance of diatoms as a major photosynthetic group and identify the most widespread and diverse genera. We also provide a new estimate of marine planktonic diatom diversity and a global view of their distribution in the world's ocean.

Author contributions: S.M., A.Z., and C.B. designed research; S.M., E.S., F.V., and J.P. performed research; S.A., J.P., P.W., and C.d.V. contributed new reagents/analytic tools; S.M., E.S., F.V., A.V., D.I., L.B., and A.Z. analyzed data; S.M. and C.B. wrote the paper; S.A. and C.d.V. provided the eukaryotic v9-18s rDNA metabarcoding dataset; and J.P. and P.W. provided sequencing of the v9-18s rDNA metabarcoding dataset.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]Present address: Biological Oceanography Division, National Institute of Oceanography, Dona Paula, Goa 403 004, India.

[2]Present address: Biological and Environmental Sciences and Engineering Division, Center for Desert Agriculture, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia.

[3]Present address: Sorbonne Universités, Université Pierre et Marie Curie (UPMC), CNRS, Institut de Biologie Paris-Seine (IBPS), Evolution Paris Seine, F-75005 Paris, France.

[4]To whom correspondence should be addressed. Email: cbowler@biologie.ens.fr.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1509523113/-/DCSupplemental.

(24), isozyme analysis (25), and microsatellite marker analysis (26) have also been used to assess diatom diversity at lower (intraspecific) taxonomic levels.

With the advent of high-throughput DNA sequencing, DNA metabarcoding has now emerged as a rapid and effective method to develop a global inventory of biodiversity that cannot be detected using classical microscopic methods (27, 28). Metabarcoding combines DNA-based identification and high-throughput DNA sequencing and is based on the premise that differences in a diagnostic DNA fragment coincide with the biological separation of species. Limitations have been identified for metabarcoding (28, 29), mainly by its dependency on PCR (and thus exposure to amplification artifacts) (30), by its susceptibility to DNA sequencing errors (31), and by the considerable investment required to build comprehensive taxonomic reference libraries (32). However, compared with previous methods, metabarcode datasets are far more comprehensive, many times quicker to produce, and less reliant on taxonomic expertise.

The choice of variable DNA regions to be barcoded needs to be evaluated carefully (33). For eukaryotes, recent reports have proposed the use of partial 18S ribosomal DNA (rDNA) sequences as potential molecular markers (34). The 18S rDNA contains nine hypervariable regions (V1–V9) (35). Amaral-Zettler et al. (34) first used the V9 region to assess general patterns in protistan diversity. They suggested that this region has the potential to assist in uncovering novel diversity in microbial eukaryotes. In the current study, we explored diatom distribution and diversity using this short (~130 base pairs) hypervariable V9 region. The availability of a taxonomically comprehensive reference database, highly conserved primer binding sites, and the potential of V9 to explore a broad range of eukaryotic diversity make this sequence well-suited as a biodiversity marker (36). We performed taxonomic profiling of 293 samples derived from 46 globally distributed sampling sites along the *Tara* Oceans circumnavigation (36–38). Experimental validation of the molecular data was established by light microscopy using samples from selected sites. Given the unprecedented genetic and geographical coverage, our study provides significant and novel insights into current patterns of diatom genetic diversity in the world's ocean.

Results

Our study, summarized in Fig. 1, was structured to develop a framework for a molecular-based analysis of marine planktonic diatom diversity, covering seven oceanographic provinces: i.e., North Atlantic Ocean (NAO), Mediterranean Sea (MS), Red Sea (RS), Indian Ocean (IO), South Atlantic Ocean (SAO), Southern Ocean (SO), and South Pacific Ocean (SPO). The metabarcoding approach we used is summarized in *SI Appendix, SI Materials and Methods* and Figs. S1 and S2. The results are presented in four broad sections: namely, (*i*) summary of the diatom metabarcoding dataset, (*ii*) local and regional novelty, (*iii*) comparison between molecular and morphological estimates, and (*iv*) global biogeographical patterns exhibited by diatoms.

Global Dataset of Diatom V9 Metabarcodes. At a cutoff level of 85% identity to sequences in our reference database (39), a total of 63,371 V9 rDNA ribotypes (represented by ~12 million sequence reads) from 293 communities could be assigned to diatoms. Rarefaction analysis indicated that these ribotypes approached saturation at a global scale (Fig. 2*A*) although individual oceanic regions, such as the NAO and RS, were far from saturation. Preston lognormal distribution extrapolated the true diatom ribotype richness to 96,710 ribotypes (fitted red curve in Fig. 2*B*), suggesting that our survey retrieved ~66% of diatom ribosomal diversity in the photic zone of the global ocean (shaded region in Fig. 2*B*). All of the ribotypes were clustered (36, 40) into biologically meaningful operational taxonomic units (OTUs), which yielded 3,875 distinct OTUs. Each OTU was represented by the most abundant ribotype in the OTU cluster. For these OTUs, Preston's veil revealed the completion in sampling to be 81.6%, with an extrapolated number of OTUs to be 4,748 (*SI Appendix*, Fig. S3).

Based on ribotype abundance, diatoms were found to be one of the most represented eukaryotic lineages [number two in eukaryotic phototrophic lineages (after the Dinophyceae, although

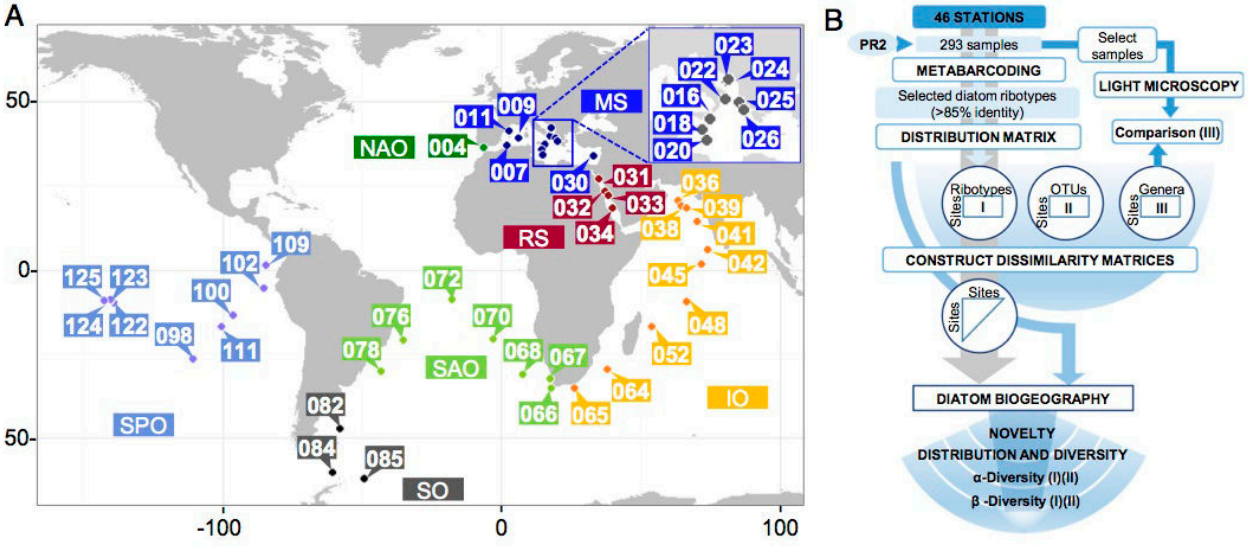


Fig. 1. Samples and methods used in the study. (*A*) Location of sampling sites (for details see ref. 37). Global diversity analysis was carried out using samples drawn from 46 global stations. At each station, the eukaryotic plankton community was sampled at two depths [subsurface (SRF) and deep chlorophyll maximum (DCM)] and fractionated into four size classes (0.8–5 μm, 5–20 μm, 20–180 μm, and 180–2,000 μm), corresponding to 293 samples altogether. IO, Indian Ocean; MS, Mediterranean Sea; NAO, North Atlantic Ocean; RS, Red Sea; SAO, South Atlantic Ocean; SO, Southern Ocean; SPO, South Pacific Ocean. (*B*) Flowchart of methods used in the study. Illumina-based sequencing was performed on each sample targeting the V9 rDNA region. All reads were quality checked and dereplicated. Taxonomy assignment was done by homology using the V9 PR2 reference database (36). From these reads, a total of 63,371 diatom-assigned ribotypes (represented by ~12 million reads) were selected for global diatom distribution and diversity analyses. Classical morphology-based identification methods using light microscopy (LM) were done on a number of selected samples to validate the molecular data.

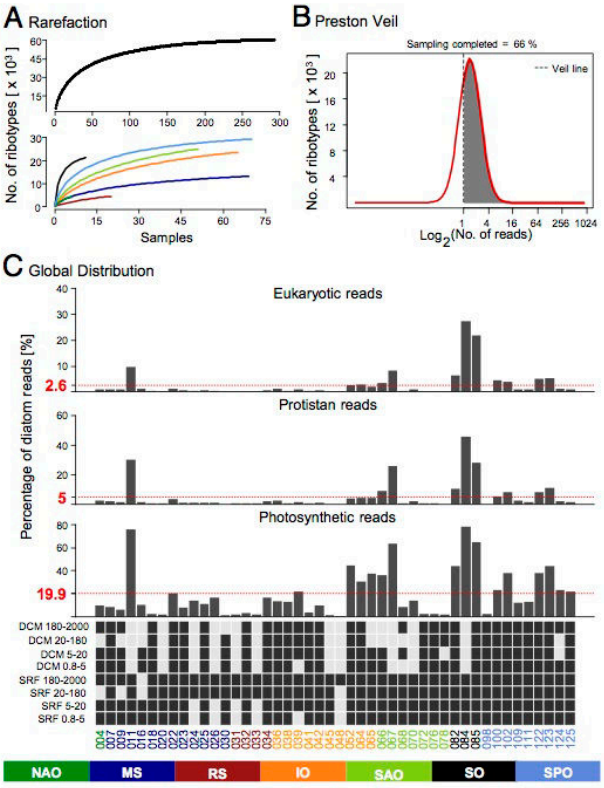


Fig. 2. Overview of the V9-rDNA diatom dataset. (*A*) V9 rDNA rarefaction curve. (*Upper*) Sample-based rarefaction curve, representing V9 rDNA richness for diatoms. (*Lower*) Each curve illustrates the estimated number of V9 rDNA sequences for each ocean province. The color code for the ocean provinces is given under the figure. Notice the scale difference in the *x* axis between *Upper* and *Lower*. (*B*) Preston log-normal distribution of diatom ribotype abundance in the entire dataset. The number of unique diatom ribotypes is plotted for logarithmically binned abundance intervals. The part of the curve on the left of Preston's Veil line (dashed black vertical line) corresponds to ribotypes with less than one read in the sample, and thus not represented in the dataset. The theoretical richness inferred from Preston's Veil was estimated to be 96,710 ribotypes, indicating 33,339 ribotypes missed during the sampling. (*C*) Percentage contribution of diatoms to the total (*i*) eukaryotic, (*ii*) protistan, and (*iii*) photosynthetic planktonic community. The red-dashed lines represent the mean percentage contribution of diatoms to each of the indicated planktonic communities. Station labels are color-coded based on the province they belong to. *Lower* shows the samples analyzed as filled boxes.

note that there are many taxa of Dinophyceae that are not photosynthetic at all or perform photosynthesis only facultatively) and number five with respect to all marine eukaryotic lineages] (36). Overall, diatom reads accounted for about 2.86% of total eukaryotic reads and 4.86% of protist ribotypes in our set of samples but represented more than 25% of the total eukaryotes at some locations: e.g., in the SO (Fig. 2*C*). Diatoms contributed ~75% to the total photosynthetic community at station 11 (MS), more than 78% and 65% at polar stations 84 and 85, respectively (SO), 44% at subpolar station 82 (SO), and more than 38% and 44% at stations 122 and 123, respectively (Marquesas Islands; tropical SPO), and globally represented 27.7% of the total eukaryotic photosynthetic planktonic community. The mean percentages of diatom reads across 46 stations were 2.6%, 5%, and 19.9% with respect to the total eukaryotic reads, protistan reads, and photosynthetic reads, respectively (Fig. 2*C*). Many tropical and subtropical stations in the MS (stations 18, 20, and 30), inner RS (stations 31, 32, and 33), IO (stations 41, 45, and 48), subtropical SAO (stations 72, 76, and 78), and in the SPO subtropical gyre (station 98) were found to be very scarce in diatom sequences in comparison with other photosynthetic groups, such as dinoflagellates and haptophytes (Fig. 2*C* and ref. 36).

Diatom Community Composition. Nearly 58% of the reads (corresponding to 33,314 ribotypes) could be assigned at least down to genus level, and the large majority (>90%) of these assigned sequences belonged to known planktonic genera (*SI Appendix*, Fig. S2). Of the 79 genera found, *Chaetoceros* was the most abundant genus, representing 23.1% of total assigned sequences. *Fragilariopsis* accounted for 15.5% of total assigned sequences, followed by *Thalassiosira* (13.7%) *Corethron* (11%), *Leptocylindrus* (10.1%), *Actinocyclus* (8.7%), *Pseudo-nitzschia* (4.4%), and *Proboscia* (3.9%) (Fig. 3, column a and Dataset S1). Only a few sequences were assigned to genera known from freshwater or benthic environments, and in most cases only with low similarity (e.g., *Fragilariforma* and *Epithemia*) (*SI Appendix*, Fig. S2), likely because of the lack of reference sequences for a number of marine planktonic genera (see *Unassigned Sequences* and *Comparison Between Light Microscopy and V9 Ribotype Counts*).

The Marine Ecosystem Biomass Data (MAREDAT) project previously provided global abundance and biomass data for all major planktonic diatoms of the global ocean ecosystem (41). Our dataset showed an overlap of 45 diatom genera with MAREDAT (*SI Appendix*, Fig. S4 *A–C*) whereas 34 genera from our study are not present in MAREDAT. A total of 23 genera present in both MAREDAT and the reference database were not found in our dataset. Most of the unmapped genera were either freshwater (e.g., *Tabellaria*, *Ulnaria*, *Urosolenia*) or benthic and marine littoral species (e.g., *Amphiprora, Caloneis*, *Ardissonea*, *Hyalodiscus*, *Pseudostriatella, Entomoneis, Phaeodactylum*), except for only a few pelagic marine genera (e.g., *Bacterosira*) (7). Some of these unmapped genera have been reported only in northern latitudes, which may explain their absence in our dataset, which is principally from the Southern Hemisphere (Fig. 1*A*). A comparison of Bacillariophyta distributions in the OBIS database (20) similarly revealed little overlap because of the lack of previous data from the locations sampled during the *Tara* Oceans expedition (*SI Appendix*, Fig. S4*D*).

Intragenus diversity was found to vary from as low as one ribotype per genus (e.g., *Nanofrustulum, Asteroplanus, Bellerochea*) to as high as 6,094 ribotypes (*Chaetoceros*) (Fig. 3, columns a and b and Dataset S1). *Chaetoceros* was found to be the most abundant and diverse genus, with 73.3% of the ribotypes (and 59.6% of the sequences) belonging to the subgenus *Phaeoceros* and the remainders to *Hyalochaetae* (Dataset S1). *Chaetoceros* (both subgenera), *Thalassiosira*, *Corethron*, and *Pseudo-nitzschia* accounted for the highest number of OTUs (Fig. 3, column c and Dataset S1). As expected, the 5- to 20-μm-size and 20- to 180-μm-size fractions contained the highest numbers of diatom ribotypes although an unexpectedly high number were also found in the smaller size fractions, belonging to smaller species (e.g., *Nanofrustulum*, *Cyclotella*, and *Minutocellus*) but also to larger species (e.g., *Attheya*, *Ditylum, and Bellerochea*) (7), perhaps derived from broken cells, broken fecal pellets, or from gametes. The 180- to 2,000-μm-size fraction contained the lowest number of ribotypes, including from chain-forming diatoms (e.g., *Hyalosira*, *Fragilaria*) and epizoic species (e.g., *Pseudohimantidium*), but also from small cells (e.g., *Nanofrustulum*), possibly having been ingested by larger organisms or otherwise associated with them or with microplastics, or retained in this fraction because of net clogging. A clear distinction was seen in the distribution among different size fractions: e.g., small and mainly solitary *Minidiscus, Attheya*, and *Minutocellus* were found highly restricted to the smallest size fractions whereas larger, chain-forming *Asterionellopsis, Lauderia*, and *Odontella* were found principally in the 20- to 180-μm-size fractions (Fig. 3, column d).

Different genera were also found to prefer different depths, such as *Actinoptychus*, *Corethron*, *Coscinodiscus*, *Fragilariopsis*, *Leptocylindrus*, and *Rhizosolenia* in subsurface (SRF) samples, whereas

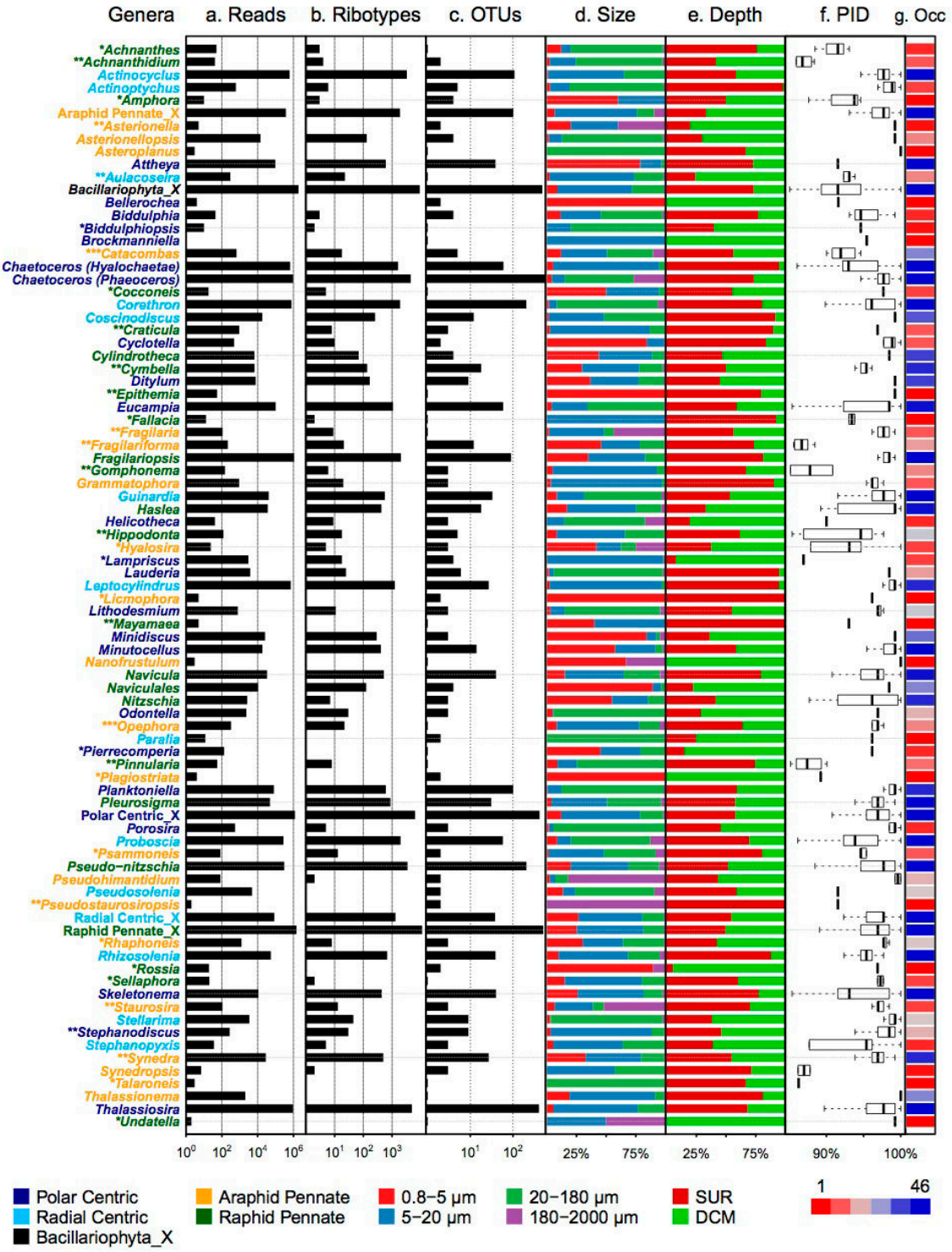


Fig. 3. Summary of diatom metabarcoding dataset. All ribotypes were clustered based on their taxonomic affiliation at the lowest taxon possible and organized under 79 genera plus five unassigned groups. The color code for a genus is as follows: dark blue, polar centric; light blue, radial centric; orange, araphid pennate; green, raphid pennate; black, unassigned Bacillariophyta. The benthic, freshwater, and brackish diatom genera are marked with *, **, and ***, respectively. (Column a) Abundances expressed as numbers of rDNA reads; (column b) richness expressed as number of unique rDNA sequences; and (column c) the corresponding number of V9 rDNA OTUs are shown for each indicated genera. (Column d) Percentage distribution of rDNA reads per size class. (Column e) Percentage distribution of rDNA reads per depth. (Column f) Boxplot showing the mean percentage sequence similarity (PID; percentage identity) to reference sequences. (Column g) Occupancy (Occ) expressed as the number of stations in which the genus was observed. The color codes for the four size classes, two depths, and occupancy are given under the figure.

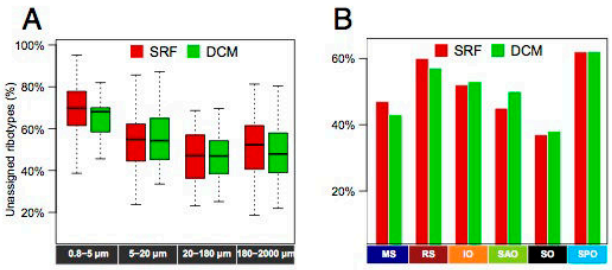


Fig. 4. Percentage of unassigned ribotypes in the *Tara* Oceans metabarcoding dataset. (*A*) Percentage of unassigned ribotypes per size class. Surface samples corresponding to 0.8–5 µm had the highest percentage of unassigned ribotypes whereas size fraction 20–180 µm had the lowest. (*B*) Percentage of unassigned diatom community at each depth in each province.

Asterionellopsis, *Bellerochea*, *Helicotheca*, *Nanofrustulum*, and *Lithodesmium* were seen mostly in deep chlorophyll maximum (DCM) samples (Fig. 3, column e). The level of percentage identity to the reference sequence also varied across genera (Fig. 3, column f). *Pseudo-nitzschia*, *Actinocyclus*, *Attheya*, *Chaetoceros*, *Eucampia*, *Fragilariopsis*, *Minutocellus*, and *Thalassiosira* were among the most cosmopolitan genera whereas many others (mainly benthic and freshwater genera) were restricted to only a few stations (Fig. 3, column g and Dataset S1).

Unassigned Sequences. We performed manual annotation on the top unassigned sequences (representing ~87% of the unassigned reads) and compared GenBank annotations with those in the PR2 database, which resulted in our being able to assign an additional 13 ribotypes (representing ~8% of the unassigned reads) from the 113 most abundant sequences to genus or species level. The best assignments and percent identity of these sequences to those present in the reference databases are shown in Dataset S2. Overall the ribotypes that could not be unambiguously assigned to any diatom genus but could be classified only as araphid or raphid pennate, polar, or radial centric, or unassigned diatom on the basis of V9 rDNA annotation (Fig. 3) represented between 31% and 81% of the total number of unique diatom ribotypes at different sampling stations (*SI Appendix*, Fig. S5). The best assignments and percent identity of these sequences to those present in the reference database are shown in Dataset S2. In general, unassigned ribotypes were particularly common in the SPO, where most of the stations are in the high nutrient low chlorophyll (HNLC) region downstream of the equatorial and Peruvian upwellings, in the IO, and in the warm and salty RS, with almost similar percentages at both depths. The diatoms in the smallest size fraction contributed most to the unknown sequences, with depth having no significant impact (Fig. 4*A*). On the other hand, the larger size fractions (20–180 µm and 180–2,000 µm) contained the lowest percentage of unassigned ribotypes, consistent with microplanktonic diatoms being the most common and the best studied. The number of unassigned sequences also varied among sampling sites, with the MS, the Benguela upwelling (station 67) (*SI Appendix*, Fig. S5), and the SO containing the best characterized diatom communities (Fig. 4*B*).

Comparison Between Light Microscopy and V9 Ribotype Counts. To investigate whether V9-based relative abundance estimates for diatoms are comparable with community composition studies based on classical morphological identification methods using light microscopy (LM), diatom counts were compared between the two methods for 15 sampling stations. A simple comparison was initially disappointing; however, the correlation between the two kinds of data was significantly improved when "unassigned" and "not known" sequences were removed from the V9 dataset and when some specific adjustments were applied (*Materials and Methods*) (Fig. 5). A few cases of mismatch still persisted: e.g., the surface sample from station 84 was dominated only by *Fragilariopsis* sp. in LM counts whereas *Chaetoceros* (*Phaeoceros* and *Hyalochaetae*) and *Fragilariopsis* were equally dominant genera along with unknown centric diatoms in the V9 dataset. However, the overall match between the two datasets was sufficiently close, thus indicating that V9 counts can provide a reliable estimate of diatom relative abundance at the genus level in a given sample.

LM also assisted in samples where we found a high percentage of unknown ribotypes. For instance, station 84 displayed abundant counts of *Asteromphalus*, a genus for which no sequences are available in the reference database. We also examined samples that contained a large number of V9 sequences that could not be assigned, specifically from stations 122–124 (*SI Appendix*, Fig. S5). In these samples, we typically observed a large number of pennate diatoms that could not be identified easily, and so we speculated that many of these unassigned sequences could be from pennate diatoms that do not yet have sequence representation in the V9 dataset. Conversely, centric genera identified by LM but not present in the V9 dataset included *Asterolampra*, *Asteromphalus*, *Climacodium*, *Dactyliosolen*, *Hemiaulus*, *Hemidiscus*, and *Lauderia*.

Global Diversity Patterns. We next examined intragenus diversity (expressed as exponentiated Shannon Diversity Index) (42) and distribution in different oceanic contexts for the 20 most abundant genera. Of these abundant genera, we found that *Pseudo-nitzschia*, *Chaetoceros* (both subgenera), and *Thalassiosira* were the most diverse genera whereas *Corethron*, *Leptocylindrus*, *Minidiscus*, and *Planktoniella* were among the least diverse and that this observation also reflected the known differences in species richness for these genera (Fig. 6*A*). Most diatom genera were seen in all oceanic provinces although their abundance patterns were highly variable: for instance, *Chaetoceros* (both subgenera), *Corethron*, and *Fragilariopsis* were highly abundant in the SO, in accordance with previous data (e.g., ref. 43); *Attheya*, *Planktoniella*, and *Haslea* were seen principally in the SPO; and *Leptocylindrus* was found to be highly abundant in the MS, especially at station 11, in line with reports from other Mediterranean sites (44). In terms of global biogeography, the diversity of each genus (expressed as the number of ribotypes) was found to be strikingly variable across the oceans (Fig. 6*B* and *SI Appendix*, Fig. S6). Three main patterns were found, with some genera having a lower diversity in the tropics (e.g., *Fragilariopsis*, *Proboscia*, and *Eucampia*), others showing lower diversity at high latitudes (e.g., *Attheya*, *Guinardia*), and others with a more uniform diversity (e.g., *Thalassosira*, possibly the most global diatom genus in our dataset). The two *Chaetoceros* subgenera showed similar distributions, with higher abundance in the SO (*SI Appendix*, Fig. S6 *B* and *C*) and high richness in coastal and open-ocean

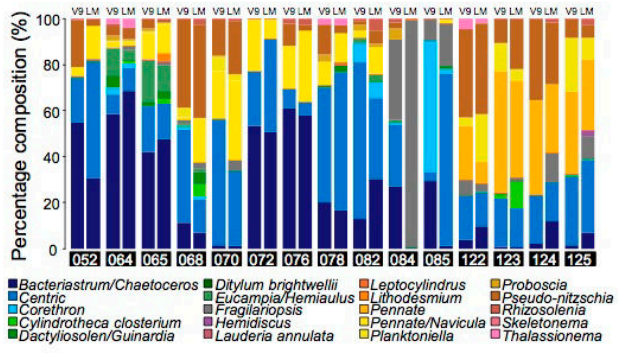


Fig. 5. Comparisons of diatom community compositions estimated from V9 rDNA counts and by light microscopy (LM). Shown are community composition profiles obtained from light microscopy and ribotype relative abundance inferred from taxonomy-based clustering of assigned ribotypes from 15 selected stations.

areas (*SI Appendix*, Fig. S6*D*). The subgenus *Phaeoceros* was more represented in the larger size fractions at almost all sites, including the offshore Atlantic, Pacific, and Mediterranean waters (Dataset S1 and *SI Appendix*, Fig. S6*C*).

Among surface samples, diversity (expressed as exponentiated Shannon Diversity Index) and evenness across oceanic provinces varied greatly, attaining the highest values in the RS, whereas, among the DCM samples, the IO showed the highest diversity; the SO was the least diverse at both depths (Fig. 7*A*). In terms of richness, the SO stations consistently showed the highest values owing to the presence of a majority of very low abundant ribotypes. Considerable variation in terms of overall ribotype diversity in different size fractions was also observed (*SI Appendix*, Fig. S7*A*). In contrast with what was observed globally for marine planktonic eukaryotes in the *Tara* Oceans dataset (36), diatom diversity did not consistently decrease with increasing size (*SI Appendix*, Fig. S7*A*). There were also no discernible differences in diatom diversity patterns between SRF and DCM samples.

Generally, the western boundary currents of the oceanic basins were the most diverse regions. Furthermore, a sudden drop in diversity was observed in the Agulhas retroflection region between the IO (station 65) and the SAO (stations 66/67/68), and from the SAO (stations 76 and 78) to the SO (stations 82/84/85) (Fig. 7*B* and *SI Appendix*, Fig. S7*B*). Diversity was significantly lower in the samples from the Maldives (station 45, North IO) but increased toward the north and the south (Fig. 7*B* and *SI Appendix*, Fig. S7*B*). Station 11 in the MS displayed the lowest diversity of all, the result of a diatom bloom that was dominated by *Leptocylindrus* (Figs. 1*C*, 6*B*, and 7*B*). In general, although the standardized abundance of diatoms showed a significant decrease from coastal to open ocean (e.g., from stations 65–67 to stations 68–78) and from surface to DCM, with the exception of the Northern IO and the SPO (Fig. 2*C*), we found no significant difference in the diversity at open ocean stations versus coastal stations (Fig. 7*C*). Indeed, diversity showed no correlation with diatom V9 sequence abundance.

We then examined whether diatom diversity follows a latitudinal gradient, as has been observed for other marine organisms (45–49). We indeed found a poleward decrease, although the trend was weak (Fig. 7*D*), most likely because of the lack of data from 50° to 60° latitudes. Analysis of the complete set of data from *Tara* Oceans will be required before drawing any concrete conclusions about latitudinal gradients.

Geographical Evenness and Community Similarity. Diatom-annotated ribotype distribution patterns were generally consistent across all of the stations, in that only a few ribotypes were abundant and the large majority of the richness was contributed by rare ribotypes (Fig. 8). The number of different ribotypes per station varied from as low as 46 (station 48; IO) to as high as 16,100 (station 85; SO), with a mean richness of 4,927. In general, it was found that the more abundant a ribotype, the more ubiquitous was its distribution (Fig. 8). Several ribotypes with considerable abundance but low occupancy were also seen, possibly indicating endemism or a marked seasonality in their occurrence (blooming species). One of the *Leptocylindrus* ribotypes was one such example. Only 23 ribotypes were found in ≥90% of the studied sites; however, they represented nearly 24% of the total relative abundance. The majority of these cosmopolitan ribotypes could not be assigned to a known diatom taxon (Fig. 8, *Lower*). A few selected unassigned ribotypes [marked with an asterisk in Fig. 8, *Lower*] were identified as *Shionodiscus bioculatus* ("*4"), *Asteromphalus* spp. ("*11"), *Pseudo-nitzschia delicatissima* ("*19") and *Thalassiothrix longissima* ("*") (*SI Appendix*, *SI Materials and Methods*). Most ribotypes with intermediate abundance aligned along a line (roughly going from occupancy: 25, evenness: 0 to occupancy: 44, evenness: 0.8), indicating a general tendency toward cosmopolitanism that is directly proportional to a deviation from an opportunistic r-strategy (corresponding to a low evenness) (50–52). Furthermore, the wide set of combinations of evenness and occupancy suggests that diatoms actually occupy all kinds of niches (*Discussion*).

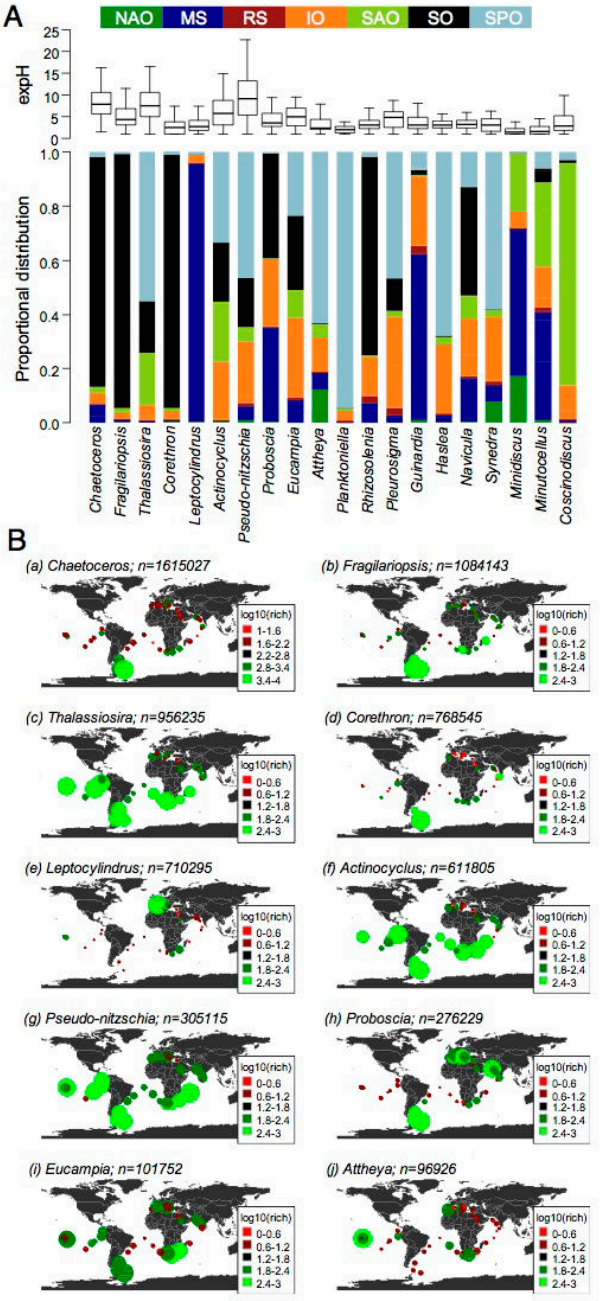


Fig. 6. Local and regional genus distribution and diversity inferred from *Tara* Oceans dataset. (*A*) Distribution of top 20 diatom genera in seven oceanic provinces. These genera accounted for 98.84% of the assigned reads in the entire dataset. (*Upper*) The variation in diversity for each indicated genus inferred from exponentiated Shannon Diversity Index (expH) across 46 stations. *Pseudo-nitzschia*, *Chaetoceros* and *Thalassiosira* were the most diverse genera whereas *Corethron* and *Minidiscus* were among the least diverse. (*Lower*) Percentage of reads in ocean provinces for the 20 most abundant genera. Bars are color-coded by ocean province, as indicated. (*B*) Global distribution and diversity of the 10 most abundant genera, which accounted for 93.3% of the assigned reads in the entire dataset. *n* is the number of reads assigned to each genus. Bubble areas are scaled to the total number of reads for each genus at each location whereas the color represents the number of unique ribotypes (red, low richness; green, high richness).

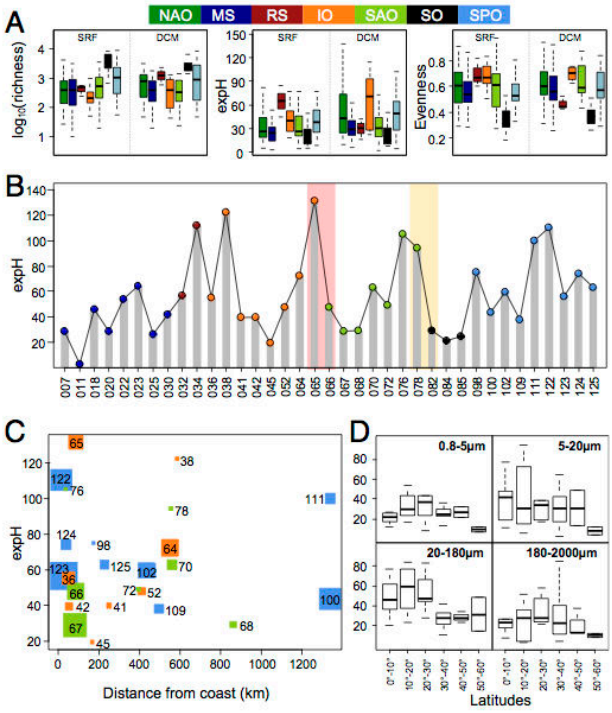


Fig. 7. Variation in diatom diversity across oceanic basins. (*A*) Variation in richness (expressed as number of unique ribotypes), diversity [expressed as exponentiated Shannon diversity index (expH)], and evenness across provinces. (*B*) Variation in diatom diversity across 37 stations (expH) for which surface samples for all size classes were available. Each station (filled circle) is color-coded based on the oceanic province it belongs to. The pink and yellow shaded regions denote the drops in diatom diversity from one province to another. (*C*) Variations in diatom diversity as a function of distance from the coast. The area of the squares represents diatom abundance (with respect to total photosynthetic reads) at each of the 37 stations analyzed. For this analysis, only stations in the major oceanic basins of the IO, SAO, and SPO were considered. (*D*) Variations in diatom diversity along absolute latitude.

The total number of ribotypes seen in the MS, RS, IO, SAO, SO, and SPO were 13,119, 4,586, 23,722, 16,269, 26,846, and 29,203, respectively. Most of the ribotypes in the SO (53.3%), SPO (33.7%), and MS (26.9%) were not found elsewhere whereas only a few ribotypes were specific to the RS (2.3%). Similarly, the IO (14.2%) and SAO (10.4%), which are transitional basins between the SPO and NAO, showed only a small number of ribotypes endemic to them (*SI Appendix*, Fig. S7 *C* and *D*). Altogether, nearly 52% (32,850 out of 63,371) of the ribotypes were seen only in one province. Interestingly, a substantial number of ribotypes were shared between two provinces [in particular, the SPO and IO (12,176 ribotypes), where the latter is downstream of the former; the SAO and SPO (9,501 ribotypes), mostly because of the coastal SAO stations; the SAO/IO (8,569 ribotypes); and the SO/IO (7,330 ribotypes)] whereas only 576 ribotypes (out of 63,371; 0.9%) were present in all oceanic provinces (*SI Appendix*, Fig. S7*D*). Diatoms thus seem to have a significant association to each oceanic basin or to basins that are physically connected (e.g., the SPO and IO via the Indonesian Passage).

The complex biogeographical patterns become clearer when considering the similarity among surface stations for which all four size fractions were available (37 stations). Stations in the SPO, SO, and MS showed the highest degree of internal similarity (Fig. 9), coherent with their relative homogeneity of conditions (for instance, the actual SPO subset is made up of tropical stations in an HNLC, iron-limited tropical region) and geographical isolation (the SO and MS). The clustering of stations revealed four major groups, including one for the MS (the most isolated case), one for the SPO, and another containing oligotrophic, seasonally stable stations where dia-

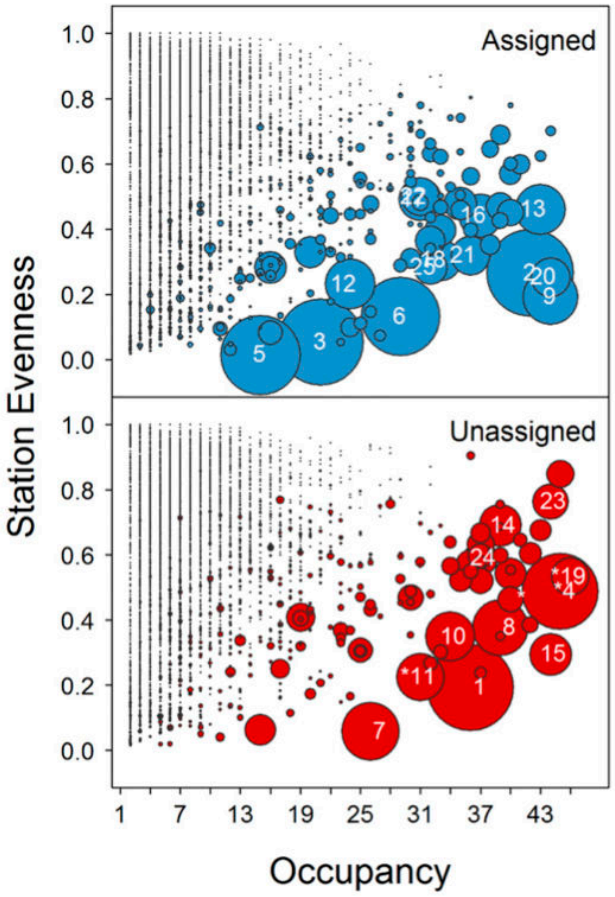


Fig. 8. Cosmopolitanism, total abundance, and station evenness of each diatom ribotype. (*Upper*) Ribotypes that could be assigned to a genus/species. (*Lower*) Ribotypes that could not be assigned to any genus. Each circle represents a ribotype (V9 rDNA), the radius being scaled to the number of reads it contains. The *x* axis corresponds to the number of stations in which a ribotype occurs; the *y* axis corresponds to the evenness of the ribotype across stations in which it occurs. The 25 most abundant ribotypes are labeled with their rank, and their assigned taxonomies are as follows: 1, Bacillariophyta_X; 2, *Fragilariopsis*; 3, *Corethron inerme*; 4, Polar Centric_X; 5, *Leptocylindrus*; 6, *Chaetoceros*; 7, *Fragilariopsis*; 8, Raphid Pennate_X; 9, *Chaetoceros*; 10, Polar Centric_X; 11, Bacillariophyta_X; 12, *Chaetoceros*; 13, *Chaetoceros rostratus*; 14, Raphid Pennate_X; 15, Araphid Pennate_X; 16, *Thalassiosira*; 17, *Thalassiosira*; 18, *Thalassiosira punctigera*; 19, Raphid Pennate_X; 20, *Thalassiosira*; 21, *Actinocyclus curvatulus*; 22, *Attheya longicornis*; 23, Bacillariophyta_X; 24, Raphid Pennate_X; 25, *Actinocyclus curvatulus*. Many ribotypes, for instance those assigned to *Leptocylindrus* (rank = 5) and *Corethron* (rank = 3), showed high abundance (larger circles), low occupancy (*x* axis), and low evenness (*y* axis). Cosmopolitan ribotypes can be identified as those with highest occupancy. A range of evenness was exhibited by them. For instance, among the most abundant sequences, ribotypes assigned to *Fragilariopsis* (rank = 2), *Chaetoceros* (rank = 9), and *Thalassiosira* (rank = 20) are cosmopolitan but with low evenness: i.e., these ribotypes are dominant only in one or two stations. Four unassigned ribotypes (*Lower*) marked with an asterisk were selected for reassignment and were identified as "*4"-*Shionodiscus bioculatus*, "*11"-*Asteromphalus* spp., "*19"-*Pseudo-nitzschia delicatissima*, and "**"- *Thalassiothrix longissima* (*SI Appendix, SI Materials and Methods*).

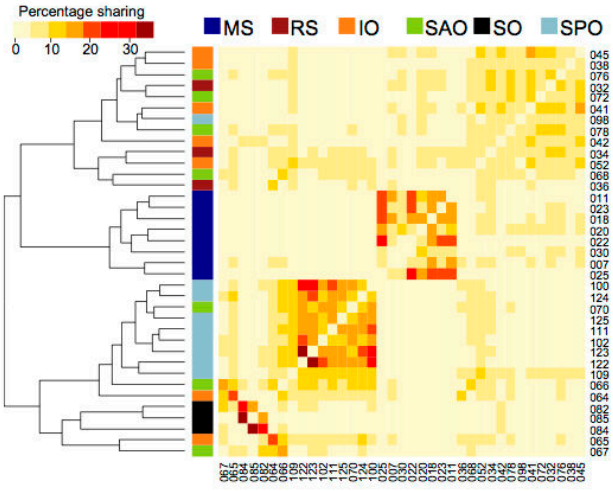


Fig. 9. Biogeographic patterns. Percentage of ribotypes shared between stations. Only those stations for which surface samples for all size classes were available are shown (37 stations). For each station, a pooled community derived from all size classes was obtained. A dendrogram of complete linkage clustering is shown. Pearson correlation was used as a distance measure to cluster stations. Two major groups were identified, one with a majority of stations from the South Atlantic Ocean, South Pacific Ocean, and Southern Ocean, and another with the Mediterranean Sea (one cluster) and low abundance stations from all oceanic regions. A substantial degree of sharing was seen among stations from the Southern Ocean, Pacific Ocean, and Mediterranean Sea. A very low internal similarity was seen in the low abundance stations.

toms were present only at low abundance. Finally, the polar SO stations and the rather coastal, mostly temperate stations around South Africa form a fourth cluster, despite their large distance and, in some cases, huge environmental gradients. This latter observation confirms that the Agulhas region, the choke point of the global circulation, is a region of intense mixing among water masses. With the exception of the low abundance clade, these clades shared a considerable percentage (~20–37%) of ribotypes within them. The community in the MS, a semienclosed basin, was most distinct from the others whereas the IO, the hub of the global surface circulation, showed the highest similarity with the others (Fig. 9 and *SI Appendix*, Figs. S7*C* and S8). The SPO and MS stations were nonetheless each seen to cluster together without any overlapping with each other, and the SO stations showed a very distinct community structure (*SI Appendix*, Fig. S8). Several specific cases illuminate the limits of this simple geographical approach and need to invoke ecological mechanisms to explain the observed patterns. For instance, station 30 in the Eastern Mediterranean Sea is part of the MS cluster even though it is in a phosphate-limited ultraoligotrophic region, unlike all of the other MS stations. Conversely, the Marquesas Islands (stations 122 and 123; SPO) are clearly under the influence of the far upstream Peruvian Upwelling (stations 100 and 102; SPO) whereas, because of the effect of a natural fertilization (53), they are quite different from the very close-by downstream stations 124 and 125. The equatorial upwelling in turn acts as a barrier, making the station further north (station 109) quite different from the others. This latter station is also upstream of all of the others (except the SO) and in fact is similar to most other stations.

Discussion

The extent of the *Tara* Oceans dataset (54) allows an unprecedented examination of the structure of plankton communities on a global scale. The current study presents an analysis of diatom community composition, based on metabarcoding using the V9 hypervariable region of 18S rDNA (36). Although this sequence has limited resolution at the species level for diatoms, we show that it is nonetheless well suited to explore genus-level diversity (*SI Appendix*, Fig. S1).

A potential caveat of metabarcoding is the presence of multiple copies of small-subunit rDNA in some species with respect to others, which is particularly pronounced in dinoflagellates (36, 55–57). Nonetheless, we argue that our diversity data for diatoms are congruent, as demonstrated by the match between molecular and morphological methods (Fig. 5). The overall coherence between these two methods indicates that rDNA copy number variation does not seem to be a major concern for diatoms (56). Conversely, the fact that the match is not perfect reveals the pros and cons of each approach. For example, LM cannot distinguish between cryptic species whereas the molecular approach cannot identify species for which there is no corresponding reference sequence. We therefore consider that the intercalibration between the two methods is very informative. Nonetheless, the diversity estimates obtained in this study should be interpreted conservatively because ribosomal diversity, rather than species diversity (58), and the fidelity of our OTU binning approach for diatoms will need to be examined with specific case studies in the future (40). A further limitation is that our dataset is based on a single sampling event at each location whereas there is known to exist substantial temporal variation in community structure (57). Our dataset therefore lacks the resolution to explore questions of endemicity.

All of the sampled communities followed comparable structural patterns, characterized by a few dominant ribotypes representing the majority of abundance and a large number of rare ribotypes. The high number of V9 reads (~1.6 million) assigned to *Chaetoceros* indicates it to be the dominant genus of marine planktonic diatoms, consistent with previous morphological surveys (e.g., refs. 59 and 60), followed by *Thalassiosira, Corethron, Fragilariopsis, Leptocylindrus,* and *Actinocyclus* (~0.5–1.0 million). The top 10 genera together accounted for more than 92.4% of the assigned reads (in terms of abundance), their dominance in the world's ocean matching findings from other studies (e.g., ref. 60). Despite their wide range, no dominant genera exhibited similar abundance and diversity patterns across stations. Among the top 10 genera, *Leptocylindrus* and *Attheya* displayed distinct geographical preferences (MS and SPO, respectively). It was observed that *Chaetoceros*, *Corethron*, and *Fragilariopsis* were more abundant in the SO, in agreement with previously reported data (61), whereas *Thalassiosira*, *Actinocyclus*, *Pseudo-nitzschia*, *Proboscia*, and *Eucampia* showed almost even worldwide distributions across all provinces (in agreement with ref. 62). In general we found complementary results when comparing genus distribution from our results (focused on the Southern Hemisphere) and previous distribution reports from the Northern Hemisphere (63). For instance, *Corethron* exhibits higher abundance in coastal locations at high latitudes in both hemispheres. These results are concordant with evidence indicating that most diatom genera are likely to be cosmopolitan due to a high chance of large scale dispersal (64). However, the diversity within each genus varied greatly across stations, suggesting shifts in community structure. Such observations warrant a more detailed analysis of the factors/processes influencing the distribution and diversity of each genus. Notably, genera that are known to be common/abundant in coastal waters were underrepresented in our dataset, like *Skeletonema, Nitzschia, Achnanthes,* and *Cocconeis*, although this finding was not observed for *Navicula* and *Pleurosigma*, which are also generally considered to be coastal genera (7).

Fourtanier and Kociolek (65) have cataloged 900 diatom genera whereas our reference database has only 159 genera (39), indicating that many genera lack sequence information. Indeed, nearly 50% of the ribotypes remain unassigned because of the lack of representatives in the reference database. It is noteworthy that one-third of the diatoms represented in the MAREDAT database do not have ribotype assignments. Moreover, different genera have different numbers of reference sequences, which may also affect the assignation of some sequences. To our

knowledge, ours is currently the largest dataset that allows assessment of the total number of marine planktonic diatom species, and our results estimate a total of 4,748 OTUs. There is nonetheless likely to be a considerable amount of novel diversity within the diatoms because many of our data are from the southern hemisphere whereas the previous studies compiled in the MAREDAT and OBIS databases have been focused largely in the North Atlantic and North Pacific (*SI Appendix*, Fig. S4). As shown in Fig. 8, we found several abundant and cosmopolitan ribotypes that were unassigned because of the lack of suitable reference sequences although more detailed sequence analysis could reveal their identity. In our opinion, it is therefore unlikely that unassigned sequences will be found to represent currently uncharacterized genera.

In general, marine planktonic diatoms are associated with nutrient-rich waters with high biomass that are commonly found in coastal waters, in upwelling areas, or during seasonal blooms in the open oceans, such as the North Atlantic spring bloom (3, 66, 67). Although our dataset contains only a few coastal sampling sites, the results reported here confirm that diatoms constitute a major component of phytoplankton and are most common in regions of high productivity (upwelling zones) and high latitudes (the Southern Ocean). However, we further show that in open ocean oligotrophic areas diatom diversity is comparable to coastal areas. At these sites, although the abundance of diatoms is low (likely because their growth is limited most of the time), they are able to survive (perhaps because of mechanisms such as dormancy, symbiosis with N-fixers, buoyancy regulation, etc.) and, for some of them, to be ready to take advantage of favorable ecological conditions as and when they arise. This reservoir of diversity is likely an essential asset ensuring an overall plasticity of response of the whole diatom community to environmental variability. The wide set of combinations of evenness and occupancy also suggests that the common view of diatoms as opportunists (i.e., r-strategists) (50–52) has to be reconsidered because they seem capable of occupying a wide range of niches and to display a diversity structure (with rare sequences being more numerous than abundant sequences) that is more akin to a gleaner (K) strategy (52). As a case in point, despite the well-known behavior of *Chaetoceros* as a local opportunist (50, 52), the impressive abundance and diversity shown here indicate that the various species do not outcompete each other. In our opinion, as a group the diatoms are therefore likely to display a continuous spectrum of different growth strategies.

Our study identified two diversity choke points for diatoms, between stations 65 and 67, and 78 and 82. These stations were situated at different sides of the Agulhas retroflection and the Drake Passage, respectively. Both areas are known to be choke points for ocean circulation (68, 69). Previous studies on diatom fossil records reported that the Agulhas choke point is not a barrier to plankton dispersal (70). However, a recent study using the entire *Tara* Oceans dataset (71) reported strong contrasts in richness across the choke point and suggested that Agulhas rings, the means of connectivity between the basins, act selectively on species distributions. Our results with diatoms are consistent with these overall patterns for the plankton community. The second choke point is constrained by the Antarctic Circumpolar Current (ACC) and is an important conduit for exchange between the Atlantic, Southern, and Pacific Oceans. At the Drake Passage, the ACC branches off to give rise to the Malvinas Current that flows northward over the Argentine slope and outer shelf, transporting saline, cold, nutrient-enriched waters (72). The high abundance of diatoms at station 82 can be attributed to these nutrient-enriched waters being transported by the Malvinas Current. A more detailed analysis of community similarity further revealed that sampling sites influenced by the ACC share similar diatom communities (Fig. 9), supporting the concept of coadapted species living within similar biomes.

The data reported here can be helpful to address Baas Becking's posit that "everything is everywhere, but the environment selects" (73). Based on Fig. 8, only a handful of diatom sequences are found everywhere (74). On the other hand, the worldwide distributions of different ribotypes from the same abundant diatom genera reported here suggest that these protists have evolved to diversify locally to varying environmental conditions to exploit a very wide range of ecological niches. This property can underpin the ecotype differentiation that has made diatoms a highly successful group of phytoplankton. Our study has laid a foundation for understanding the processes that constrain marine diatom communities and that control their biodiversity, and the extensive physical, chemical, and other contextual data collected during the *Tara* Oceans expedition (37, 54) should allow a wide range of ecological and evolutionary questions to be addressed.

Materials and Methods

Diatom Metabarcoding Dataset. For the present study, 293 global samples encompassing 46 stations from the photic zone [subsurface (SRF) and deep chlorophyll maximum (DCM)] were used that corresponded to four size classes (0.8–5 μm, 5–20 μm, 20–180 μm, and 180–2,000 μm). A total of 63,371 V9 rDNA diatom-assigned ribotypes (represented by ~12.4 million reads) were retrieved from the 293 communities. Please see de Vargas et al. (36) for details on the sequencing and taxonomic assignation of the V9 sequences used in this study.

Taxonomy-Based Clustering. Metabarcodes were clustered based on their taxonomic affiliation at the level of genus and were organized under 86 genera. Five additional unassigned classes (unassigned, unassigned polar centric, unassigned radial centric, unassigned raphid pennate, and unassigned araphid pennate) were defined to accommodate those reference sequences ($n = 416$) for which genus assignment was not available. Genus distribution and diversity were assessed for most represented genera.

Global Distribution Analysis. Deviations from Preston's log-normal distribution were used to estimate the completeness of richness sampled. Also, the information from the samples was used to extrapolate the number of ribotypes that might be found if sampling were more intensive. The relation between abundance, occurrence, and evenness of each ribotype was assessed. Pielou's evenness (75) and the exponentiated Shannon–Weiner H′ diversity index (42) were used as estimates of diversity. The percentage of shared ribotypes was calculated for each pair of stations, and a Spearman correlation was used as a distance measure to cluster stations. Compositional similarity between stations was computed based on a Hellinger-transformed abundance matrix and incidence matrix using Bray–Curtis and Jaccard indices, respectively, as a measure of β-diversity. Nonmetric multidimensional scaling was performed to visualize the level of similarity between different stations. For all statistical analyses, a value of $P < 0.05$ was considered significant. All of the data analyses were performed in R (76).

ACKNOWLEDGMENTS. We thank Achal Rastogi, Yann Thomas, and Marie-José Garet-Delmas for technical support. We thank the commitment of the following people and sponsors who made the *Tara* Oceans Expedition 2009–2013 possible: Centre National de la Recherche Scientifique and the Groupement de Recherche GDR3280, European Molecular Biology Laboratory, Génoscope/Commissariat à l'Energie Atomique, the French Government "Investissements d'Avenir" programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), MEMO LIFE (ANR-10-LABX-54), Paris Sciences et Lettres (PSL*) Research University (ANR-11-IDEX-0001-02), the Agence Nationale de la Recherche (ANR) projects FRANCE GÉNOMIQUE (ANR-10-INBS-09-08), POSEIDON (ANR-09-BLAN-0348), PROMETHEUS (ANR-09-GENM-031) and PHYTBACK (ANR-2010-1709-01), European Union Framework Programme 7 (MicroB3/No.287589), European Research Council Advanced Grant Award (to C.B.) (Diatomite: 294823), Agnès b., the Veolia Environment Foundation, Region Bretagne, World Courier, Illumina, Cap L'Orient, the Eléctricité de France (EDF) Foundation EDF Diversiterre, Fondation pour la Recherche sur la Biodiversité, the Prince Albert II de Monaco Foundation, Etienne Bourgois, and the *Tara* schooner and its captain and crew. E.S. was partially supported by a grant from the Ministero dell'Istruzione dell'Università e della Ricerca RITMARE project. *Tara* Oceans would not exist without continuous support from 23 institutes (oceans.taraexpeditions.org). This article is contribution 36 of *Tara* Oceans.

1. Smetacek V (1998) Diatoms and the silicate factor. *Nature* 391:224–225.
2. Falkowski PG (2002) The ocean's invisible forest. *Sci Am* 287(2):54–61.
3. Armbrust EV (2009) The life of diatoms in the world's oceans. *Nature* 459(7244): 185–192.
4. Nelson DM, Treguer P, Brzezinski MA, Leynaert A, Queguiner 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.

5. Field CB, Behrenfeld MJ, Randerson JT, Falkowski P (1998) Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* 281(5374): 237–240.
6. Falkowski PG, Barber RT, Smetacek V (1998) Biogeochemical controls and feedbacks on ocean primary production. *Science* 281(5374):200–207.
7. Round FE, Crawford RM, Mann DG (1990) *The Diatoms: Biology and Morphology of the Genera* (Cambridge Univ Press, Cambridge, UK).
8. Kooistra WHCF, Gersonde R, Medlin LK, Mann DG (2007) The origin and evolution of the diatoms: Their adaptation to a planktonic existence. *Evolution of Primary Producers in the Sea*, eds Falkowski PG, Knoll AH (Elsevier, Boston), pp 207–249.
9. Bowler C, Vardi A, Allen AE (2010) Oceanographic and biogeochemical insights from diatom genomes. *Annu Rev Mar Sci* 2:333–365.
10. Smetacek V (2012) Making sense of ocean biota: How evolution and biodiversity of land organisms differ from that of the plankton. *J Biosci* 37(4):589–607.
11. Tréguer PJ, De La Rocha CL (2013) The world ocean silica cycle. *Annu Rev Mar Sci* 5:477–501.
12. Sournia A, Chrdtiennot-Dinet MJ, Ricard M (1991) Marine phytoplankton: How many species in the world ocean? *J Plankton Res* 13(5):1093–1099.
13. Mann DG, Droop SJM (1996) Biodiversity, biogeography and conservation of diatoms. *Hydrobiologia* 336:19–32.
14. Guiry MD (2012) How many species of algae are there? *J Phycol* 48:1057–1063.
15. Mann DG, Vanormelingen P (2013) An inordinate fondness? The number, distributions, and origins of diatom species. *J Eukaryot Microbiol* 60(4):414–420.
16. Lundholm N, et al. (2006) Inter- and intraspecific variation of the *Pseudo-nitzschia delicatissima*-complex (Bacillariophyceae) illustrated by rRNA probes, morphological data and phylogenetic analyses. *J Phycol* 42:464–481.
17. Behnke A, Friedl T, Chepurnov VA, Mann DG (2004) Reproductive compatibility and rDNA sequence analyses in the *Sellaphora pupula* species complex (Bacillariophyceae). *J Phycol* 40:193–208.
18. Degerlund M, Huseby S, Zingone A, Sarno D, Landfald B (2012) Functional diversity in cryptic species of *Chaetoceros socialis* Lauder (Bacillariophyceae). *J Plankton Res* 34: 416–431.
19. Hasle GR, Syvertsen EE (1996) Marine diatoms. *Identifying Marine Diatoms and Dinoflagellates*, ed Tomas CR (Academic, San Diego), pp 5–385.
20. OBIS (2015) Data from the Ocean Biogeographic Information System. Intergovernmental Oceanographic Commission of UNESCO. Available at www.iobis.org. Accessed July 29, 2015.
21. Logares R, et al. (2014) Patterns of rare and abundant marine microbial eukaryotes. *Curr Biol* 24(8):813–821.
22. Beszteri B, John U, Medlin LK (2007) An assessment of cryptic genetic diversity within the *Cyclotella meneghiniana* species complex (Bacillariophyta) based on nuclear and plastid genes, and amplified fragment length polymorphisms. *Eur J Phycol* 42(1): 47–60.
23. Gallagher JC (1980) Population genetics of *Skeletonema costatum* (Bacillariophyceae) in Narragansett bay. *J Phycol* 16:464–474.
24. Rynearson TA, Armbrust EV (2000) DNA fingerprinting reveals extensive genetic diversity in a field population of the centric diatom *Ditylum brightwellii*. *Limnol Oceanogr* 45:1329–1340.
25. Skov J, Lundholm N, Pocklington R, Rosendahl S, Moestrup O (1997) Studies on the marine planktonic diatom *Pseudo-nitzschia*. 1. Isozyme variation among isolates of *P. pseudodelicatissima* during a bloom in Danish coastal waters. *Phycologia* 36:374–380.
26. Evans KM, Hayes PK (2004) Microsatellite markers for the cosmopolitan marine diatom *Pseudo-nitzschia pungens*. *Mol Ecol Notes* 4:125–126.
27. Yu DW, et al. (2012) Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods Ecol Evol* 3:613–623.
28. Bittner L, et al. (2013) Diversity patterns of uncultured Haptophytes unravelled by pyrosequencing in Naples Bay. *Mol Ecol* 22(1):87–101.
29. Will KW, Mishler BD, Wheeler QD (2005) The perils of DNA barcoding and the need for integrative taxonomy. *Syst Biol* 54(5):844–851.
30. Bellemain E, et al. (2010) ITS as an environmental DNA barcode for fungi: An *in silico* approach reveals potential PCR biases. *BMC Microbiol* 10:189.
31. Coissac E, Riaz T, Puillandre N (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. *Mol Ecol* 21(8):1834–1847.
32. Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol Ecol* 21(8): 2045–2050.
33. Riaz T, et al. (2011) ecoPrimers: Inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Res* 39(21):e145.
34. Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM (2009) A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS One* 4(7):e6372.
35. Ki JS, Han MS (2005) Molecular analysis of complete SSU to LSU rDNA sequence in the harmful dinoflagellate *Alexandrium tamarense* (Korean isolate, HY970328M). *Ocean Sci J* 40:155–166.
36. de Vargas C, et al.; Tara Oceans Coordinators (2015) Ocean plankton: Eukaryotic plankton diversity in the sunlit ocean. *Science* 348(6237):1261605.
37. Pesant S, et al.; Tara Oceans Consortium Coordinators (2015) Open science resources for the discovery and analysis of *Tara* Oceans data. *Sci Data* 2:150023.
38. Karsenti E, et al.; Tara Oceans Consortium (2011) A holistic approach to marine ecosystems biology. *PLoS Biol* 9(10):e1001177.
39. Guillou L, et al. (2013) The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* 41(Database issue, D1):D597–D604.
40. Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M (2014) Swarm: Robust and fast clustering method for amplicon-based studies. *PeerJ* 2:e593.
41. Leblanc K, et al. (2012) A global diatom database: Abundance, biovolume and biomass in the world ocean. *Earth Syst Sci Data* 4:149–165.
42. Hill MO (1973) Diversity and evenness: A unifying notation and its consequences. *Ecology* 54:427–432.
43. Gersonde R, Zielinski U (2000) The reconstruction of late quaternary antarctic sea-ice distribution: The use of diatoms as a proxy for sea-ice. *Palaeogeogr Palaeoclimatol Palaeoecol* 162(3-4):263–286.
44. Siokou-Frangou I, et al. (2010) Plankton in the open Mediterranean Sea: A review. *Biogeosciences* 7(5):1543–1586.
45. Tittensor DP, et al. (2010) Global patterns and predictors of marine biodiversity across taxa. *Nature* 466(7310):1098–1101.
46. Fuhrman JA, et al. (2008) A latitudinal diversity gradient in planktonic marine bacteria. *Proc Natl Acad Sci USA* 105(22):7774–7778.
47. Sul WJ, Oliver TA, Ducklow HW, Amaral-Zettler LA, Sogin ML (2013) Marine bacteria exhibit a bipolar distribution. *Proc Natl Acad Sci USA* 110(6):2342–2347.
48. Vyverman W, et al. (2007) Historical processes constrain patterns in global diatom diversity. *Ecology* 88(8):1924–1931.
49. Rodríguez-Ramos T, Marañón E, Cermeño P (2015) Marine nano- and microphytoplankton diversity: Redrawing global patterns from sampling-standardized data. *Glob Ecol Biogeogr* 24:527–538.
50. Reynolds CS (2006) *The Ecology of Phytoplankton* (Cambridge Univ Press, Cambridge, UK).
51. Margalef R (1978) Life forms of phytoplankton as survival alternatives in an unstable environment. *Oceanol Acta* 1:493–509.
52. Barton AD, Dutkiewicz S, Flierl G, Bragg J, Follows MJ (2010) Patterns of diversity in marine phytoplankton. *Science* 327(5972):1509–1511.
53. Blain S, Bonnet S, Guieu C (2008) Dissolved iron distribution in the tropical and sub tropical South Eastern Pacific. *Biogeosciences* 5:269–280.
54. Bork P, et al. (2015) Tara Oceans. Tara Oceans studies plankton at planetary scale: Introduction. *Science* 348(6237):873.
55. Galluzzi L, et al. (2004) Development of a real-time PCR assay for rapid detection and quantification of *Alexandrium minutum* (a Dinoflagellate). *Appl Environ Microbiol* 70(2):1199–1206.
56. Godhe A, et al. (2008) Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. *Appl Environ Microbiol* 74(23): 7174–7182.
57. Nolte V, et al. (2010) Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. *Mol Ecol* 19(14):2908–2915.
58. Piganeau G, Eyre-Walker A, Grimsley N, Moreau H (2012) How and why DNA barcodes underestimate the diversity of microbial eukaryotes. *PLoS ONE* 7(4):10.1371.
59. VanLandingham SL (1968) *Catalogue of the Fossil and Recent Genera and Species of Diatoms and Their Synonyms. Part II. Bacteriastrum Through Coscinodiscus* (Verlag von J. Cramer, Lehre, Germany), pp 494–1086.
60. Hinder SL, et al. (2012) Changes in marine dinoflagellate and diatom abundance under climate change. *Nat Clim Chang* 2:271–275.
61. Smol JP, Stoermer EF (2010) *The Diatoms: Applications for Environmental and Earth Sciences* (Cambridge Univ Press, Cambridge, UK).
62. Chamnansinp A, Li Y, Lundholm N, Moestrup Ø (2013) Global diversity of two widespread, colony-forming diatoms of the marine plankton, *Chaetoceros socialis* (syn. *C. radians*) and *Chaetoceros gelidus sp. nov. J Phycol* 49:1128–1141.
63. GBIF (2014) Updated GBIF Work Programme 2014–2016 (Global Biodiversity Information Facility, Copenhagen), Version 2015. Available at www.gbif.org/. Accessed February 12, 2016.
64. Vanormelingen P, Verleyen E, Vyverman W (2008) The diversity and distribution of diatoms: From cosmopolitanism to narrow endemism. *Biodivers Conserv* 17:393–405.
65. Fourtanier E, Kociolek JP (2003) Catalogue of the diatom genera (vol 14, pg 190, 1999). *Diatom Res* 18:245–258.
66. Cervato C, Burckle L (2003) Pattern of first and last appearance in diatoms: Oceanic circulation and the position of polar fronts during the Cenozoic. *Paleoceanography* 18:1055.
67. Bopp L, Aumont O, Cadule P, Alvain S, Gehlen M (2005) Response of diatoms distribution to global warming and potential implications: A global model study. *Geophys Res Lett* 32:1–4.
68. Cunningham SA, Alderson SG, King BA, Brandon MA (2003) Transport and variability of the Antarctic Circumpolar Current in Drake Passage. *J Geophys Res* 108:8084.
69. Siedler G, Griffies S, Gould J, Church J (2013) *Ocean Circulation and Climate: A 21st Century Perspective* (Academic, Oxford).
70. Cermeño P, Falkowski PG (2009) Controls on diatom biogeography in the ocean. *Science* 325(5947):1539–1541.
71. Villar E, et al.; Tara Oceans Coordinators (2015) Ocean plankton: Environmental characteristics of Agulhas rings affect interocean plankton transport. *Science* 348(6237): 1261447.
72. Peterson RJ, Stramma L (1991) Upper level circulation in the South Atlantic Ocean. *Prog Oceanogr* 26(1):1–73.
73. Baas Becking LGM (1934) *Geobiologie of Inleiding tot de Milieukunde* (W.P. Van Stockum & Zoon, The Hague, The Netherlands) (in Dutch).
74. Medlin LK (2007) If everything is everywhere, do they share a common gene pool? *Gene* 406(1-2):180–183.
75. Pielou E (1966) The measurement of diversity in different types of biological collections. *J Theor Biol* 13:131–144.
76. R Development Core Team (2009) R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna).

Supporting Information

Insights into global diatom distribution and diversity in the world's ocean

Shruti Malviya, Eleonora Scalco, Stéphane Audic, Flora Vincent, Alaguraj Veluchamy, Julie Poulain, Patrick Wincker, Daniele Iudicone, Colomban de Vargas, Lucie Bittner, Adriana Zingone, Chris Bowler

This file includes:

- Materials and Methods
- Figures S1 to S8
- Dataset 1 and 2

SI Materials and Methods

Distance based Analysis

The PR2 database v99 (1) contains 2,947 full length 18S unique diatom sequences. These sequences were aligned and sequence variations along the entire sequence were used to define the hypervariable regions. Entropy calculation was done on all reference sequences (Fig. S1A). Pairwise distances were calculated for the full length and all hypervariable regions using Kimura-2-parameter model (2). V4 and V9 sequences were used to check the performance in differentiating the four prominent phylogenetic clades of diatoms, i.e., radial centric, polar centric, araphid pennate and raphid pennate. Each of the V4 and V9 hypervariable regions and full-length 18S rDNA sequences were aligned using MUSCLE and phylogenetic inference was done with NJ algorithm using pairwise distances in MEGA6 (2). The tree was statistically tested using 1000 bootstraps. A reference database was obtained and all the reference sequences were aligned. Shorter sequences (less than 125 nucleotides) along with extremities were eliminated to obtain the same sequence lengths. To evaluate the ability of the V9 region to differentiate between the intragenus and intergenus variation among diatom V9 sequences, we calculated p-distance between all pairs of reference sequences (Fig. S1).

Extracting diatom V9 metabarcodes from global eukaryotic protistan metabarcoding data set

A total of ~580 million quality-checked reads, representing ~2.3 million unique rDNA ribotypes (V9 region of 18S rDNA), were generated from 334 photic-zone plankton communities sampled during the *Tara* Oceans expedition (3). The *Tara* Oceans expedition (4-5) covered seven oceanographic provinces, i.e., North Atlantic Ocean (NAO), Mediterranean Sea (MS), Red Sea (RS), Indian Ocean (IO), South Atlantic Ocean (SAO), Southern Ocean (SO), and South Pacific Ocean (SPO). At each station, plankton communities were obtained for four size fractions from two water-column depths (SRF and DCM). Total nucleic acids (DNA + RNA) were extracted from all samples, and the hyper-variable V9 region of the nuclear 18S rDNA was PCR-amplified (3). The V9 reads were quality checked and to reduce the influence of PCR and sequencing errors, only sequences seen in at least two different samples with at least 3 copies were retained. The sequences have been deposited in GenBank (see [4], Accession: PRJEB4352; ID: 213098). Taxonomy assignments for all ribotypes were obtained through annotation against an expert-curated V9 reference database (for details, see [3]) using the global alignment search strategy implemented in the ggsearch36 program (*Fasta* package). This reference database contains sequences from both cultured strains and the environment, and contained 1,232 unique diatom V9 reference sequences corresponding to 159 genera, with most genera being represented by more than one sequence (Fig. S2). Of the 159 genera in the reference data set, we retrieved 87 genera in our data set. However, only 79 out of 87 were assigned at an identity greater than 85% and were selected for further analysis.

These unique barcodes were taxonomically assigned to known eukaryotic entities based on the PR2 database (1). From this, metabarcodes assigned to diatoms, at a percentage identity of ≥ 85% to the reference sequence, were selected. When BLAST results gave rise to more than one unique best hit, a last common taxonomy of the best BLAST hits was created [3]. Moreover, in order to improve the assignation of the barcodes that couldn't be assigned to the genus level, as the PR2 database version we used was based on a former release of Genbank [3], PR2 assignations were also compared to Genbank assignations (release 210 from October 2015). We manually checked each assignation, and kept the best of two (PR2 or Genbank assignation) based on the best percentage identity value and BLAST scores. We could thus improve our conclusions and assignations, in particular when working with sequences that could not be assigned to the genus level (SI Dataset 2).

All the barcodes were clustered into biologically meaningful operational taxonomic units (OTUs) using the 'Swarm' approach (6). This method uses 1 base pair difference (local threshold) between barcodes. It also overcomes input-order dependency induced by centroid selection, a typical bias of classical clustering methods (6).

Morphological analyses

The 20-180 µm size fraction samples selected for microscopy analyses included SRF and DCM samples from the Cape Agulhas region (Stations 52, 64, 65, 66, 67, and 68), the South Atlantic transect (Stations 70, 72, 76, and 78), the Southern Ocean stations (Stations 82, 84 and 85), and South Pacific Ocean stations (Stations 122, 123, 124, and 125). Three ml of each sample was placed in an Utermöhl chamber with a drop of calcofluor dye (1:100,000), which stains cellulose thus allowing to better detect and identify diatom species. Cells falling in 2 or 4 transects of the chamber were identified and enumerated. Phytoplankton species were identified and enumerated using a light inverted microscopy (Carl Zeiss Axiophot200) at 400x magnification. The identification was performed at the species level when possible.

Reassignment of unknown diatom ribotypes

Four diatom V9 rDNA ribotypes were chosen (marked with asterisk (*) in Fig. 8, *lower panel*) for reassignment, based on their presence in the top 20 most abundant unassigned diatom ribotypes in the whole *Tara* metabarcode dataset. The goal was to amplify a longer 18s rDNA fragment of the target diatom from 18s rDNA preamplified samples in order to improve the quality of the sequence taxonomy. Preamplification of 18s rDNA was performed on DNA extraction of ethanol fixed sea water collected for each of the *Tara* metabarcoding samples (TV9_172, TV9_225, TV9_361 and TV9_339). DNA was extracted with MasterPureTM DNA/RNA purification kit (Epicenter) and PCR amplified using the universal-eukaryotic primers (forward Euk-A [5'-aacctggttgatcctgccagt-3'] and reverse Euk-B [5'-tgatcctcctgcaggttcacctac-3']) from Medlin et al. (7). Amplifications were performed with the Phusion™ high-fidelity DNA polymerase (Finnzymes) in a 50-µL reaction volume, using the following PCR parameters: 30 s at 98 °C; followed by 15 cycles of 10 s denaturation at 98 °C, 30 s annealing at 57.5 °C, and 30 s extension at 72 °C; with a final elongation step of 10 min at 72 °C. PCR product was purified with Nucleospin® PCR Clean-up (Macherey-Nagel).

For each target diatom ribotype, the equivalent 18s rDNA preamplified sample in which its relative abundance was the highest was chosen for PCR (Polymerase Chain Reaction), in order to maximise chances of amplifying the ribotype with highly specific reverse primers.

The forward primer chosen was the D512F (D512F: 5'-ccgcgtaattccagctccaatagcg-3') universal

diatom primer from Zimmerman et al. (8). The reverse primer was designed in order to find the 3' end consecutive eight base pairs 100% specific to the target sequence that matched the lowest number of non-specific sequences in the sample. Four ribotypes and their respective reverse primer sequences are listed below:

Sample ID	Ribotype md5sum ID	Reverse primer sequences
TV9_172	01eb4d181204cc0e142f55f1632b0b8c	172_rev: 5'-aggttcggacaagttctcgcggtcag-3'
TV9_225	4d2d2df1f3cdb2080ace0b23c17928be	225_rev: 5'-ttcctactaaatgataaggtttagacgagt -3'
TV9_361	ba6c7a54f4f24e0888797d4e062cda61	361_rev: 5'-ggggacaagttctcgcggctaacat-3'
TV9_339	8e6521a0e8234f3660e5c0d302c33da9	339_rev: 5'-gcggagacaagttctcgcgacagat-3'

TV9_179: St82\0.8-inf\srf; TV9_225: St85\20-180\srf; TV9_361: St123\5-20\srf; TV9_339: S122\5-20\dcm

The unassigned V9 sequence was cut in windows of 8 base pairs and each of them was mapped against all positions of the OTU sequences present in the sample under Perl 5 (version 16, subversion 2 (v5.16.2)). A heatmap of hits was obtained, giving the number of times each window perfectly matched a position in the V9 sequences of the sample. The best primer candidates were then extended to 26 base pairs on average, and the final primer was chosen based on its position in the target sequence, close to the end of the V9, its GC content, Tm and checked on IDTDNA Oligo Analyzer 3.1 (http://eu.idtdna.com/calc/analyzer). Temperature gradient PCR from 58 to 68 °C were performed to obtain highest specificity of the primers to the target DNA.

Amplifications were performed with the Phusion™ High-Fidelity DNA Polymerase (Thermo-Scientific™) in a 20-µL reaction volume, using the following PCR parameters: 30 s at 98 °C; followed by 33 cycles of 10 s denaturation at 98 °C, 30 s annealing from 66 to 68 degrees, and 60 s extension at 72 °C; with a final elongation step of 10 min at 72 °C. DNA was extracted from agarose gel with Nucleospin® Gel and PCR Clean-up (Macherey-Nagel) and directly sent to GATC Biotech for paired-end Sanger Sequencing. Resulting sequences were assigned by blastn in NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

References

1. Guillou L, et al. (2013) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* 41(D1):D597-D604
2. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol Biol Evol* 30:2725-2729.
3. de Vargas C, et al. (2015) Eukaryotic plankton diversity in the sunlit global ocean. *Science* 348(6237):1261605.
4. Pesant S, et al. (2015) Open science resources for the discovery and analysis of Tara Oceans data. *Sci Data* 2:150023.
5. Karsenti E, et al. (2011) A holistic approach to marine eco-systems biology. *PLoS Biol* 9:e1001177.
6. Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M (2014) Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* 2:e593.
7. Medlin L, Elwood HJ, Stickel S, Sogin ML (1988) The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. *Gene* 71:491-499.
8. Zimmerman J, Jahn R, Gemeinholzer B (2011) Barcoding diatoms: evaluation of the V4

142 subregion on the 18S rRNA gene, including new primers and protocols. *Org Divers Evol*
143 11(3):173-192.
144 9. Leblanc K, et al. (2012) A global diatom database - abundance, biovolume and biomass in
145 the world ocean. *Earth Syst Sci Data* 4:149-165.
146 10. OBIS (2015) Data from the Ocean Biogeographic Information System. Intergovernmental
147 Oceanographic Commission of UNESCO. Web. http://www.iobis.org (consulted on
148 2015/07/29).
149

Chapter 3: Global scale patterns of diatom interactions in the open ocean

Summary

3.1. Investigating microbial interactions at large spatial scale : introduction

Our insight about global diatom and marine microbes biodiversity has considerably moved forward with advances in sequencing as the ocean contains an immense diversity of marine microbes. Different functional groups of bacteria, archaea and protists arise from this diversity to dominate various habitats and drive globally important biogeochemical cycles (Menden-Deuer et al., 2016). The study of their distribution and associated activities often focused on resource availability and abiotic conditions, but those factors are insufficient to explain major patterns of functional group dominance in the sea (Green et al., 2008). The continual reshaping of communities by mortality, allelopathy, symbiosis and other processes show that such community interactions exert strong selective pressure on marine microbes (Strom, 2008). This reflects the "Eltonian shortfall", introduced by Hortal et al., 2015 in his review on current major flaws in biodiversity research and refers to our lack of knowledge about "biotic interactions" (see Annex B) among species, and among groups of species especially in the marine microbial world. However, because interactions can affect population dynamics, it is expected that the signatures of microbial interactions are imprinted in microbial survey datasets.

The study of microbes associations in other systems has demonstrated that they are essential for community stability - such as a healthy microbiota - and that their disturbance such as the overgrowth of a competitive pathogenic species can induce dysbioses (microbial imbalance) and diseases (Silverman et al., 2010; Frank et al., 2011). Understanding how marine microbial interactions, and that of diatoms in particular, structure the planktonic community is of key importance in a changing ocean. Microbial interactions have been increasingly investigated using "co-occurrence networks", an approach presented below.

3.1.1. Biotic interaction and species co-occurrence

The impact of biotic associations on the distribution and diversity of organisms is a topic that has been debated intensively. The degree to which non-climatic factors could shape the

distribution of species has been discussed for near a century (Wallace, 1878; Baselga et al., 2012). Specifically, there is interest in understanding the extent to which occurrences of species are constrained by the distributions of other species, at broad scales of resolution and extent (Gravel et al., 2011). Empirical studies have historically focused on competition (Gause, 1934; Hardin, 1960) showing that in its extreme form, competition leads to co-exclusion of the interacting species (MacArthur, 1972).

The earliest significant work on the relationship between biotic interactions and community assembly was developed by Diamond (Diamond 1975), that he formalized as the "Assembly Rules". These rules suggested that competitive exclusion, and not dispersal, was responsible for the avian assemblages in Guinea that seemed like a "checkerboard". Connor and Simberloff later attacked this theory in 1979, arguing that his patterns were not tested for significance, i.e., that they were not compared to what Diamond could have obtained at random. A form of "null hypothesis" was proposed in 2001 by Hubbell: he advanced that abundance and diversity of a community are driven by random dispersal, speciation, and extinction, a framework known as the unified neutral theory of biodiversity (Hubbell, 2001). As reviewed in Faust and Raes, 2012, the application of niche and neutral theory based analysis to microbial distributions supported both models as determining factors of species composition. A recent study provided the first comprehensive simulation of the expected co-occurrence between two species arising from all possible combinations of direct biotic interaction types (Araújo et al., 2014). The study shows that similar co-occurrences can be achieved by different interactions, leading to the conclusion that co-occurrences alone are not sufficient to provide insight into the biotic interactions generating them. Similarly, it has been argued that pairwise interactions between species cannot provide general principles about the dynamics and organization of complex communities at global scale (McGill et al., 2006) (**Figure 3.1**).

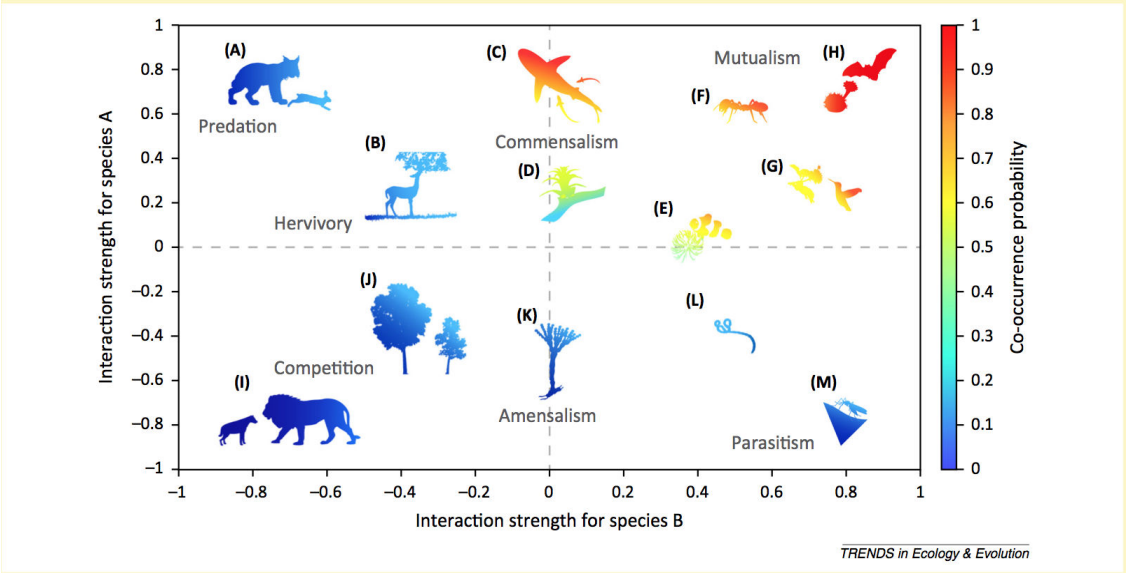


Figure 3.1. Interaction strength and probability of co occurrence.
Co-occurrence probability between two different species across biotic interaction space. Biotic interaction space is a representation of all possible types of interactions across two axes, each indicating the direction and the strength of the interaction for each species. Examples are for (A) predation of *Oryctolagus cuniculus* by *Lynx pardinus*, (B) hervibory by *Odocoileus virginianus*, commensalism by (C) *Remora brachyptera* and *Carcharhinus melanopterus*, (D) epiphytic bromeliad (fam. Bromeliaceae), and (E) *Amphiprion percula* and *Entacmea quadricolor*; examples of mutualism for (F) shelter-defense interaction between *Pseudomyrmex ferruginea* and *Cecropia peltata*, of (G) pollination of *Helicornia caribaea* by *Eulampis jugularis* and of (H) pollination of *Stenocereus thurberi* by *Leptonycteris curasoae*; competition between (I) *Panthera leo* and *Crocuta crocuta* and between (J) *Swietenia mahagoni* individuals; amensalism produced by (K) *Penicillium expansum*, and parasitism of (L) virus of genus Ebolavirus and (M) *Anopheles gambiae* mosquito, which is itself host for *Plasmodium falciparum*. (Morales-Castilla et al., 2015)

However, there is today experimental evidence that biotic interactions affect species range (Araújo et al., 2014; Bateman et al., 2012), inducing non-random co-distribution of species at large spatial scales of hundreds of kilometers for macro-organisms (Gotelli et al., 2010), both at regional and continental scale. Due to sequencing technologies, our knowledge about the composition of microbial communities from diverse environments has greatly expanded through alpha and beta analysis of diversity. It is time to investigate community structure through the characterization of inter-taxa interactions, by exploring multi species-specific microbial associations in a holistic manner. Co-occurrence networks are becoming a useful tool based on the rationale that community structure is driven, amongst other things, by ecological interactions between species and therefore that the non-random patterns of species distribution can be used to infer these interactions.

3.1.2. Co-occurrence inference with microbial survey data

The ultimate goal of microbial associations inference in large spatial metabarcoding datasets is to find which pairs of species co-occur more than what would be expected at random, or on the contrary if two species tend not to co-occur in the same samples. The main starting point underlying co-occurrence patterns prediction is the existence of a community matrix, where sites are in columns and species are in rows. The matrix is filled by abundance or presence/absence data. The similarity between the distribution of any two species is quantified using different types of measures (**Figure 3.2**). The data is then compared to a set of randomized matrixes, in order to detect non-random co-occurrence patterns and assess the significance of the measures (Faust and Raes, 2012).

Here, I will focus only on methodologies and challenges faced by inference methods related to predicting pairwise relationships between two species (correlative techniques) as implemented in the CoNet software used in this thesis, and leave out those that can predict more complex ones such as regression-based interactions (**Figure 3.2**).

Choosing the correlation measure. When performing pairwise comparison between species abundances to assess their distribution similarity, the choice of the measure is crucial, in that each measure captures different trends and this is often neglected. Pearson correlation captures linear dependencies; Spearman correlations will detect monotonic relationships through rank ordering. More exotic measures such as KullBack-Leibler divergence can be used that calculates information gap, or the Bray-Curtis score that measures dissimilarity rather than similarity. However, methods such as association rule mining or regression analysis can also help reveal complex associations, though being more computationally intensive. In order to maximize chances of detecting pairwise interactions, it is generally advised to combine measures and keep the statistically significant interactions that meet a consensus.

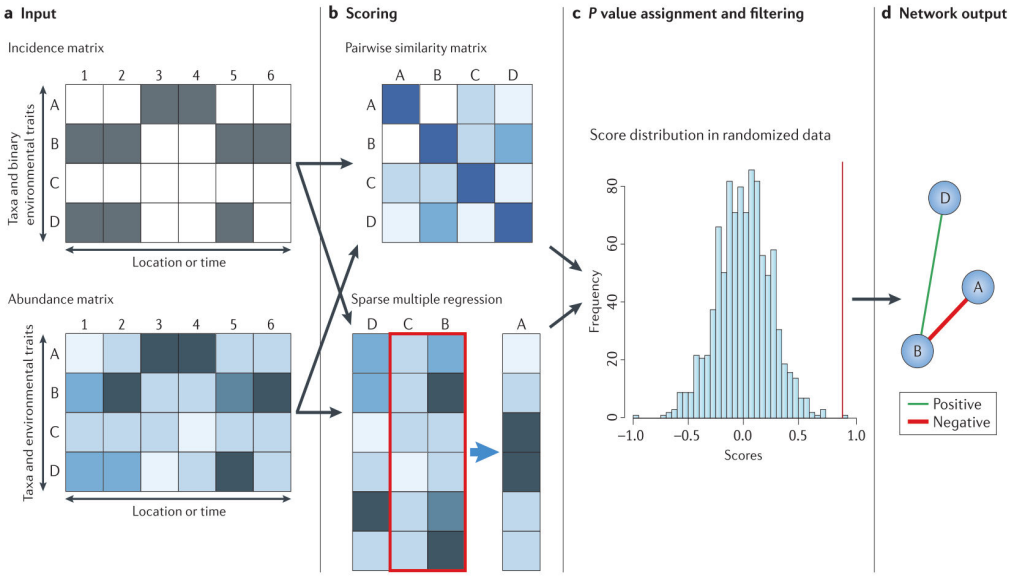


Figure 3.2. Principle of similarity- and regression-based network inference.
The goal of network inference is to identify combinations of microorganisms that show significant co-presence or mutual exclusion patterns across samples and to combine them into a network. a | Network inference starts from an incidence or an abundance matrix, both of which store observations across different samples, locations or time points. b | Pairwise scores between taxa are then computed using a suitable similarity or distance measure). In contrast to similarity-based approaches, multiple regression can detect relationships that involve more than two taxa. c | In the next step, a random score distribution is generated by repeating the scoring step a large number of times (often 1,000 times or more). The random score distribution computes the P value (that is, the probability of obtaining a score by chance that is equal to or better than the observed score) to measure the significance of the predicted relationship. The P value is usually adjusted for multiple testing with procedures such as Bonferroni or Benjamini–Hochberg.d | Taxon pairs with P values below the threshold are visualized as a network, where nodes represent taxa and edges represent the significant relationships between them. The edge thickness can reflect the strength of the relationship (Faust and Raes, 2012).

Data sparseness. When a species is displayed as "0" in the abundance matrix, it can either be attributed to the physical absence of a species (called structural 0) or insufficient sequencing depth to capture the species (sampling 0). This is currently a bottleneck in microbial co-occurrence inference, as "absence of presence" is hard to prove. This is particularly problematic for any analysis based on log transformation and presence/absence, as scientists need to add a pseudo-count and thus assume that all species are present. Moreover, data sparseness can cause what is known as the "double 0 problem": many species are rare in large-scale metagenomic samplings, and will display

many “0” across the data that will induce a spurious positive correlation between the two organisms. The Bray-Curtis dissimilarity, which is robust to spurious correlations from presence-absence data, can be used. Keeping organisms present in at least 25% of the samples can lower down the impact of 0’s (**Figure 3.3**).

Compositional effects. Due to unequal sampling effort or unequal sequencing depth, working with absolute read counts in high throughput sequencing is rare, following which scientists generally normalize their data and obtain relative abundance data. As the total abundance across all the samples is often constrained by a constant sum (1, if the abundance is divided by the total abundance in each sample), an increase in the relative abundance of one species will induce the decrease of the other, and result in a negative correlation and falsely predict negative interactions (**Table 3.1**).

Table 3.1.	Site 1	Site 2	Site 3
Species 1	30 (0,3%)	30 (0.15%)	30 (0.1%)
Species 2	70 (0.7%)	170 (0.85%)	270 (0.9%)

Table 3.1. Compositional bias
Cell value corresponds to absolute number of species counts; relative abundance is indicated in parenthesis. The abundance of species 1 is constant, the abundance of species 2 increases. Even though absolute number of species 1&2 are not correlated, when normalized they are negatively correlated. This is a spurious correlation.

This negative bias is known as the compositional bias (Aitchison, 1981) and is common in spatial metagenomic surveys, because abundance of species is usually very uneven. Dealing with compositionality using a permutation procedure is one solution that is detailed below, in order to discard any interactions that are due to skewed distributions.

Assessing significance of a score. Once the initial correlation score matrix is computed, several procedures can be carried out to ensure statistical robustness. One can compute the **confidence interval** around initial score with bootstrapping. By sub sampling the initial abundance matrix (with replacement) 1,000 times, and recomputing all pairwise correlation measures, this step aims to create a confidence interval around the initial scores. All co-

occurrences not within the limits of the 95% confidence interval are discarded. Once a correlation score is obtained, it is necessary to assess its significance with respect to a null distribution by calculating the p-value - the probability of finding the observed or more extreme result under the null hypothesis. A **random score** distribution is generated by repeating the scoring step a large number of times (often 1,000 times or more), through shuffling and renormalisation of the matrix. This null distribution represents the distribution of scores if the two organisms were distributed at random. In CoNet, a final specific p-value, per method and per pairwise correlation is computed as the probability of the null value (the mean of the null distribution) under the bootstrap distribution; the corresponding interaction is considered significant if the p-value < 0.05. Finally, when performing a large number of tests (i.e., a large number of comparisons), as is the case when we perform pairwise comparisons between one organism and all the others in an abundance matrix, the probability of finding a significant test by chance alone increases (also known as false positives, or Type I error). **Multiple testing corrections** control the number of false positive interactions, and produces an adjusted p-value threshold for each interaction whose significance under the null hypothesis can then be assessed. The p-value is usually adjusted for multiple testing with procedures such as Bonferroni or Benjamini–Hochberg.

Confounding factors and indirect dependencies. Co-occurrence analysis reflects the sum of all possible factors affecting species distribution, both biotic (other species), and abiotic. Sometimes, two species can be positively correlated because both of them are negatively correlated to a third environmental parameter. Another confounding factor is the potential effect of other species on a particular pairwise association (Morueta-Holme et al., 2015). However, tools such as interaction information (Meyer et al., 2008), and network deconvolution (Feizi et al., 2013) can help detecting indirect dependencies (**Figure 3.3**).

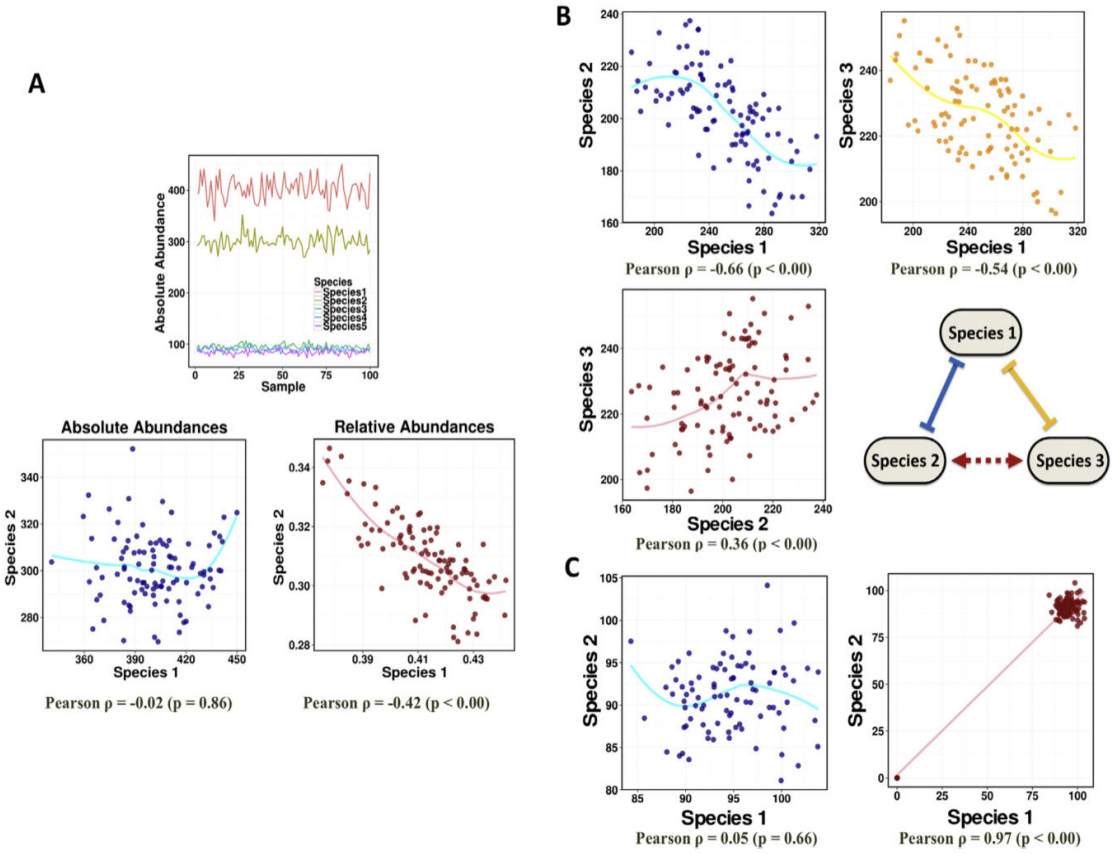


Figure 3.3. Challenges in using correlations from metagenomic survey data to infer microbial interactions.
(A) Compositional Effect: In a community with five species, where species 1 and species 2 have uncorrelated absolute abundances, their abundances appear correlated after being normalized into relative abundances. (B) Indirect Correlations: The abundances of species 2 and species 3 are positively correlated not because the two species interact with each other, but because they both interact with species 1 and are negatively correlated with it. (C) The abundances of species 1 and species 2 are not correlated. However, if there are sites where neither of the species is present, the two species can have an observed positive correlation (Li et al., 2016).

The final output of microbial co-occurrence inference is a table, or a matrix that should contain the following information: identification of significant pairwise correlations, the corresponding scores, and associated p-value. However, when co-occurrence analysis is performed on large amounts of data, extracting the information requires appropriate tools to visualize, and analyze the resulting graph.

3.1.3. The contribution of graph theory

"Good information design reveals the greatest number of ideas in the shortest time, with the least ink, in the smallest amount of space" Edward Tufte, 1983, The Visual Display of Quantitative Information.

Networks in biology emerged from the need to investigate a system not only as an individual component, but as a whole. From protein-protein interactions to metabolic pathways or transcriptional regulations, networks have flooded biology. The best way to represent all the elements of a system, and their interactions, is to draw the connections between them, forming a graph. Graph theory can measure modularity, connectance, degree distributions, and help connect elements of graph structure with notions of functions in the system. Each graph G, is formed from nodes, the objects, connected by edges, the relationships. We often say that the graph G = (N, E) is composed of N nodes, and E edges (Newman, 2003). Many descriptors and network topologies can be mathematically derived from the structure and shape of the graph, a discipline known as graph theory. As soon at the graph represents a system with defined objects, it becomes a network.

Graphs come in many different shapes, connected or not, weighted or not, directed or undirected (**Figure 3.4**). A connected graph is when all nodes can be reached by following a path, meaning a succession of connected nodes. A graph is not connected if a node, or a cluster of nodes, can't be reached by following a path; we say that the graph has more than one connected component. The number of connected components in a network is an indicator of the global connectivity of the network so that a low number of connected components relates to strong network connectivity, because many nodes are connected. In a weighted graph, the relationships between nodes can be quantitatively characterized by their weight, for example, the number of collaborations between two research institutes. Finally, directed or undirected graphs represent existence or not of a direction between two nodes, for example a trophic cascade between a predator (node 1), that eats a prey (node 2) and the relation is always in the same direction.

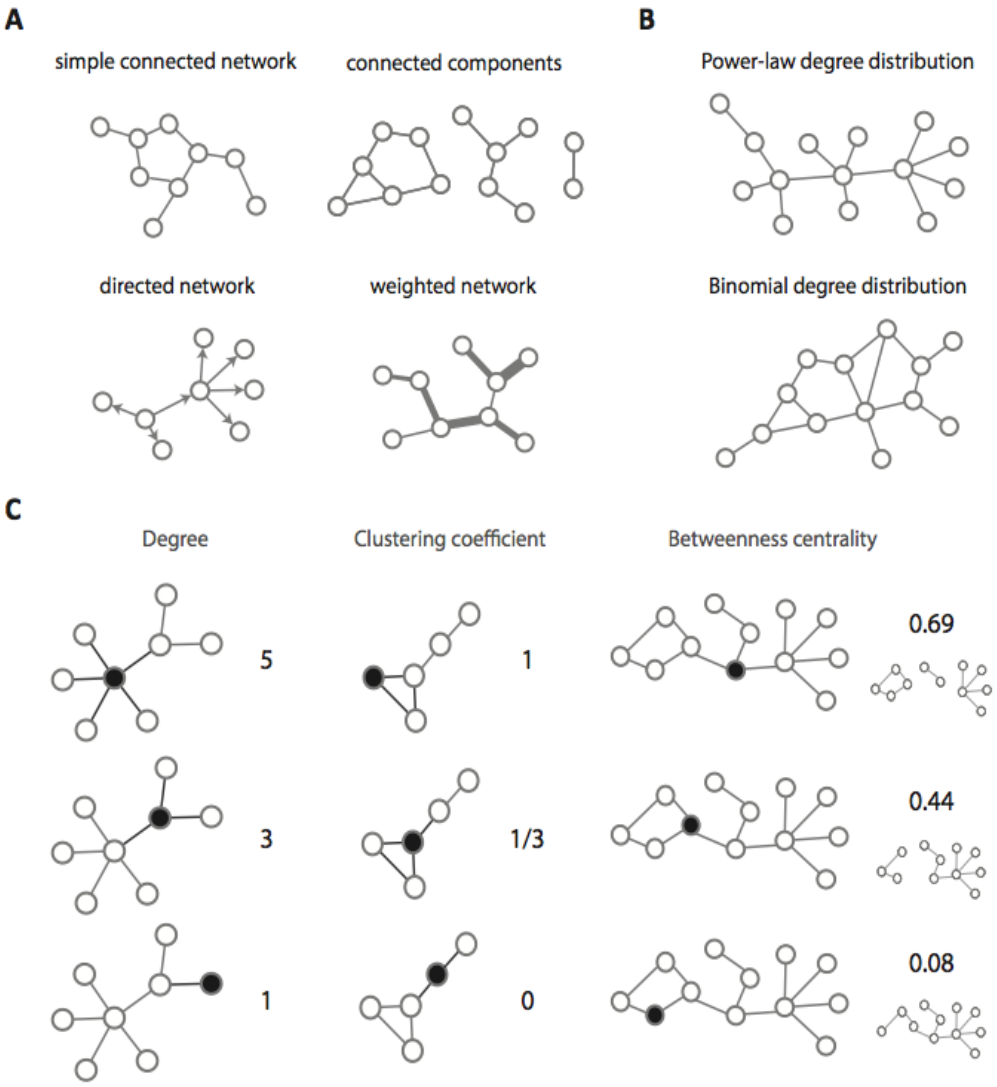


Figure 3.4. An overview of graph types and graph theory metrics.
Graphs are composed of nodes and edges, here represented as circles and links between circles, respectively. A) Graphs can be simple, directed, weighted, completely connected or composed of connected components. B) Two graph topologies that differ in their degree distribution are shown: a power law degree distribution and the characteristic bino- mial degree distribution of randomly generated graphs. C) The topology of nodes is characterized by graph theory metrics including degree, clustering coefficient and centrality measures such as betweenness centrality. The number next to each graph corresponds to the metric value of the coloured-in node (Perez, 2015).

Several properties and parameters can be derived from a network. Parameters such as the number of nodes, and the number of edges, will give estimates of the graph's size. Two graphs, with different numbers of nodes or edges can furthermore be compared using other

classes of measures such as connectance, which measures the proportion of realized edges out of the total possible combinations.

(1.1) $$\text{Connectance} = \frac{|E|}{|N|(|N|-1)/2}$$

With $|E|$ the number of edges in the graph and $|N|$ the number of nodes.

In general, graph descriptors can be divided in two categories, the ones that measure properties of individual nodes and edges, and the ones that evaluate global properties of the graph. **Node Degree** is the number of edges connected to a node and can be used to classify nodes by their connectivity $\underline{d}$:

(1.2) $$\underline{d} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} e_{ij}}{|N|} = \frac{|E|}{|N|}$$

where $|N|$ is the number of nodes, $|E|$ is the number of edges, and the edge e_{ij}= 1 if the i^{th} and j^{th} nodes are connected, otherwise e_{ij} = 0 (Newman, 2003). Once the node degree of each node is computed, a common analysis is the node degree distribution, that sometimes follows the characteristic power law distribution, meaning that the graph has a very limited number of highly connected nodes (called hubs) and a majority of nodes that are lowly connected.

Clustering coefficients. On a node basis, it is also possible to measure local connective behavior of each node, also known as the clustering coefficient of the node (Newman, 2003).

(1.3) $$c_i = \frac{\sum_{i}^{N} \sum_{j}^{N} e_{jk} * e_{ij} * ejk}{d_i(d_i-1)/2}$$

Where the numerator of the fraction is the number of triangles (a set of three nodes connected in triangle) through node i, with $e_{jk} * e_{ij} * ejk = 1$ if i, j , and k are all connected. d_i is the degree of the node. The clustering coefficient (always between 0 and 1) of a node expresses the connectivity between neighbours: if all its neighbours are connected then a node has a clustering coefficient of 1. The global clustering coefficient of a graph is the average of the nodes' clustering coefficients. The average clustering coefficient distribution gives the average of the clustering coefficients for all nodes n with d = 2, 3, ... degrees. In particular, the average clustering coefficient distribution was used to identify the modular organization of metabolic networks (Ravasz et al., 2002).

Centrality measures. Centrality measures are used to evaluate the position of a node within a graph, to determine its centrality with respect to other nodes in the graph. For instance, betweenness centrality will evaluate the position of a node globally, by evaluating the importance of that node relative to all paths in the graph, and reflects its importance in the overall structure of the network.

(1.4) $$bc_i = \sum_{j,k,j\neq k}^{N} \frac{p_{jk}(i)}{p_{jk}}$$

Where $p_{jk}(i)$ is the number of paths between node j and k that go through i while p_{jk} is the total number of paths going through j and k (Newman, 2003). The betweenness centrality of each node is a number between 0 and 1 and reflects the amount of control that this node exerts over the interactions of other nodes in the network. The centrality measure probably reflects the importance of a node in maintaining the overall structure of a network (Newman, 2003). Though the most appropriate centrality measure to use is not always straightforward (Albert et al., 2000; Iyer et al., 2013).

All the above measures characterize global topological properties of a graph and its nodes. However, **modules** can be used to reveal sub global topology by finding structurally meaningful subgraphs, which can be interpreted as a form of topological clustering on the graph (Newman, 2006). A module is defined as a subgraph whose connectivity pattern

between its members is greater than the connectivity patterns with nodes outside that subgraph. The right algorithm to partition a graph in modules depends on the context and the system studied. They have been useful in protein protein networks, to detect groups of proteins that perform specific biological functions (Li et al., 2008) or in food webs to find functional modules such as trophic networks (Brilli et al., 2010; Pascual et al., 2005).

Ecological quantitative methods have been used since a long time to understand ecosystems, and their relations to their abiotic environment; seemingly, graph theory is now being applied to microbial co-occurrence to evaluate a community's inter-connected structure, and its relation to its environment. For example, the Human Microbiome Project has collected 1200 samples across different body sites, providing relative abundance for more than 40,000 taxonomic groups in total through a spatial metabarcoding sampling (Qin et al., 2010). This data has been used to infer interactions between species based on co-occurrence networks, revealing strong niche speciation, with most microbial associations occurring within body sites, and a number of accompanying inter body site relationships (Faust et al., 2012). Later, analysis of gut microbial data revealed that healthy subjects have a more robust network than diseased subjects (Naqvi et al., 2010). In another ecosystem, the soil, robustness simulations of networks were conducted by removing OTUs with decreasing centrality, to identify key microbial genera from natural forest and agricultural plantations (Steele et al., 2011; Lupatini et al., 2014). The combination of graph theory and microbial association networks has increased in the past few years and can reveal meaningful correlations between taxa and environmental conditions (Barberán et al., 2012; De Menezes et al., 2015).

Networks are not the only means of visualisation of inferred co-occurrences. Adjacency matrices are the linear, algebraic formulation but also a visual representation of graphs, where nodes are represented in rows and columns, while edges are encoded as entries in the matrix. While this has been an appropriate and efficient way to analyse graphs until now, it scales poorly to large networks and is not suitable to look at individual node topologies such as the clustering coefficient that detects potential keystone species.

3.1.4. Interpretation of microbial co-occurrence networks

The structural properties of microbial co-occurrence networks have been characterized to infer biological attributes of the community, such as resilience and disturbance, but despite the promises and power of graph theory, finding the relevant methods to discern patterns with such complex assemblages is challenging. From structure of co-occurrence networks to systems properties, the following section investigates some current research directions in the field of microbial co-occurrence inference.

- **Interpreting co-occurrence networks**

Comparing microbial co-occurrence networks with social, ecological, and protein protein networks. Steele has applied network analysis based on natural environmental co-occurrence patterns to examine the more complex interactions amongst microbes (Steele et al., 2011) by including marine bacteria, archaea, protists and environmental parameters. The obtained network was compared with a random network (equal number of edges and nodes) generated based on the Erdös-Réyni model using the Random Network plugin in Cytoscape (Shannon et al., 2003). Subnetworks are created with taxonomic assignment rather than module detection and high positive correlation values are interpreted as potential strong direct dependencies such as symbiosis or parasitism. By calculating the clustering coefficient, the characteristic path length, and the node degree distribution, and comparing these descriptors with a random network but also to non-biological networks, they show that the microbial association network has small world properties (i.e., that nodes are more connected than a random network of similar size), and is more highly correlated than at random. The fitting of a truncated power law function, to their degree connectivity distribution, yielding similar parameter values to other ecological networks (and smaller than that of social or protein interaction networks), is an argument to claim the observation of “meaningful, non-random relationships over time”.

Linking keystone species with betweenness centrality and network robustness. In 2014, Lupatini et al., investigated the soil microbiome in Brazilian biomes using correlation network analysis and high throughput sequencing (HTS). Co-occurrence is inferred based on

OTUs grouped by genus (i.e., all OTUs assigned to the same genus are clustered into one, and the abundance is the sum of all those OTUs), arguing that this prevents potential taxonomic misclassification due to sequencing bias, which is debatable. The only score used is the Pearson correlation, no p-value correction for multiple testing was performed, no importance of indirect association was assessed. The resulting network was compared to 1000 random networks using the Erdös-Renyi model to compare the structures, based on average clustering coefficient, average path length, and modularity and assign p-value to topological measures obtained from the original dataset. Significant p-values justified that the original network was non-random, however no clear biological interpretation of the network descriptors is provided. Betweenness centrality, the fraction of shortest paths going through a given taxon to another, is proposed as a proxy for keystone species.

One option to identify potential keystone OTUs is network robustness analysis that consists in iteratively removing nodes and evaluating the structure of the network resulting from the removal, especially secondary extinctions (Pascual et al., 2005). In the gut microbiome, the data has revealed that healthy subjects have more robust networks than diseased subjects (i.e., the removal of particular nodes had less impact on the overall structure) (Naqvi et al., 2010).

Linking specialists and generalist organisms with network structure. Soil microbial communities were investigated over large spatial scales by Barberan et al., 2012. Pyrosequencing of the 16S rRNA gene followed by filtering and Spearman's rank correlation were used to infer networks. Regarding network topology descriptions, average node connectivity, average path length (value of the average distance between all pairs of node: 5.53 edges), diameter (value of the longest distance: 18 edges), cumulative degree distribution, clustering coefficient (value of how nodes are embedded in their neighborhood and thus, the degree to which they tend to cluster together: 0.33), and modularity (value: 0.77; above 0.4, it is considered that networks form modules) were calculated. The authors applied the terms generalist and specialist to OTUs that are more or less present in the samples. This can be confusing in the ecological interpretation of co-occurrence network, where specialists and generalists rather refer to whom OTUs are connected to, i.e., connected to a restricted or broad number of different species. Nonetheless, networks of

generalist species were less connected and more compartmentalized than specialist networks, attributed to the higher habitat variability covered by generalist network.

Size fractionated sampling reveal differential connectivity patterns. Milici et al., 2016 explored spatial co-occurrences of bacterioplankton taxa in the Atlantic Ocean using state-of-the-art co-occurrence inferences (SparCC). Three different size fractioned networks were computed, corresponding to Free Living, Small Particle Associated, and Large Particle Associated communities across 27 stations and five different depths, and are further compared with respect to the absolute number of correlations in the different networks. Results suggest high connectivity of Free Living bacterioplankton, primarily based on the comparison of positive correlation numbers and node degree.

- **Validating the network and framework foundations**

It is obvious that correlation does not imply true biotic interactions. Positive associations can reflect parasitism, cross-feeding, symbiosis, shared preference for a similar niche. Similarly, negative associations can reflect competition, allelopathy, predator - prey interactions or an opposite preference for a specific niche. Several proposals have been made in order to validate co-occurrence networks like the use of simulation, amongst which the possibility to model microbial populations using simple rules about their growth (Lotka-Volterra for example) in order to simulate dynamics of complex multispecies assemblages. This will define the range of parameters, equations, and situations in which a robust inference about biotic interactions can be made. Cazelles et al., 2015, performed multispecies simulations and discussed whether the inference of the structure of interaction networks is feasible from co-occurrence data. This is illustrated through modeling the considerable variety of mechanisms, causing pairwise associations and reveals how hard is can be to infer species interactions from co-occurrence analysis.

Testing co-occurrence inference using Lotka-Volterra dynamics. Berry et al., in 2014 made a good step forward by simulating multispecies microbial communities with known interaction patterns using the generalized Lotka-Volterra dynamics. Co-occurrence networks

were built using different association metrics and evaluated to see how well the networks revealed the underlying interactions, and how experimental and ecological parameters could affect network inference and interpretation. The result show that co-occurrence works under certain conditions, but that they lose interpretability (i.e., decreased specificity) when the effects of habitat filtering become significant. Topological features are associated with keystone species such as betweenness centrality. Through simulations, Berry was able to propose conditions increasing reliability of inference: extensive sampling breadth, collecting intrinsic ecological parameters such as diversity, appropriate association metric used. He showed that networks computed with a low number of sites are susceptible to false positives and that samples should originate from similar environments. He further articulated best practices that are coherent in light of the current methodological challenges: filter out infrequent species and try to have a least 20% similarity between samples and high species richness, have a high coverage sequencing, include as many samples as possible (at least 25), include samples from similar environments to avoid habitat filtering, use absolute abundance if possible, or correct for compositionality, and check indirect correlations.

Developing a framework with alternative null models. In 2016, Morueta-Holme et al., proposed a few interpretations based on simulated interactions at local scale. For example, the degree of each species characterizes the number of its association partners that are either positive or negative associations. The ratio between both can measure the species role, as an attractive aggregator, or a repulsive segregator. The modularity represents the overall structure of the network, indicating the amount of division of the network into clusters that are more densely connected to each other. The modules correspond to subsets of species more likely to be mutually associated. Their probability of co-occurring across local communities is higher as compared with the probability of co-occurring with other species of the whole community. These modules could represent guilds, trophic motifs, co-evolved organisms and species. An interesting input from their work is the use of alternative explicit regional null models, taking into account processes such as dispersal and broad scale climate patterns.

Mining the scientific literature for available interactions was advanced as an option to test true positives (Li et al., 2016) as was used for the study of Protein Protein Interactions (Cohen, 2005). Each cited paper above should however be interpreted in light of software biases exposed by Weiss et al., 2016, where the authors compare popular microbial association tools (LSA, CoNet, SparCC and simple Pearson and Spearman) on mock communities.

The arrival of graph theory and graphical visualization was key for the field of microbial ecology, because pairwise correlations in tabular format were impossible to interpret. Due to the lack of appropriate framework and technical issues, very few community scale network studies have been performed in marine microbial ecology, however this is not true for macroecology, in which the field of Ecological Networks could be of extreme use for us.

3.1.5. Inspiration from ecological networks and insight from the macro world

Investigating the role of species associations through forms of networks, and their impact on the community structure started early in ecology. Before the microbial and HTS era, biotic interactions were restricted to the macro realm. Data was collected from small-scale field observations and/or manipulative experimentation, generally using macro-organisms such as plant-pollinators, food webs, or host parasites associations.

Back in 1880, Lorenzo Camerano was the first one to depict food webs in a diagrammatic manner (**Figure 3.5**), in the manuscript “On the equilibrium of living beings by means of reciprocal destruction.” At the time, the scientific community struggled to know which species were beneficial or harmful for agriculture. In response to this binary way of thinking, Camerano responded: “To have an exact and clear idea, I repeat, of the relations between, for example, insectivorous birds and phytophagous insects, and between these and plants, these groups cannot be studied separately. Rather it is necessary to study each in relation to all other animals to see the general laws governing the equilibrium of animal and plant species.”

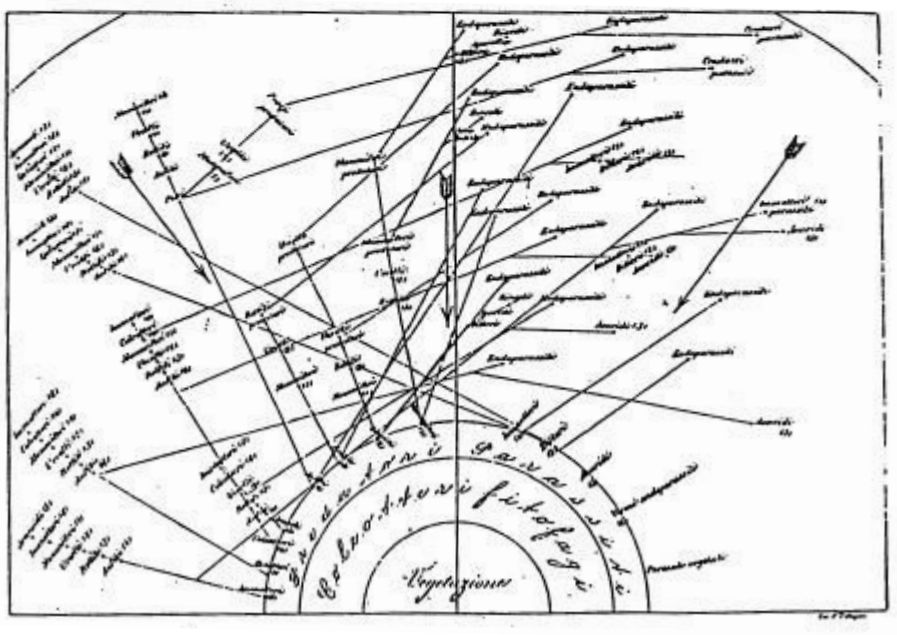


Figure 3.5. Camerano's food web (plate 2).
Web of interactions between Coleoptera, enemies of Coleoptera, and ecological enemies of those enemies.

In practice these studies are restricted to studying a few species on a local scale, for collecting data is time consuming, and impossible to use in order to understand small scale associations across large regions. But, along with Darwin's first observation of the "tangled bank", this sparked the concept of the "Web of Life" that represents the global interdependence among species without neglecting species interactions. Network approaches in ecological research emphasize the patterns of interactions among species and have proved their usefulness in order to establish frameworks and to ask questions about ecology and evolution of observed patterns (Bascompte, 2009).

The vast majority of all ecological networks are centered on a focused, single type of interaction creating bipartite networks: predator-prey (Pimm, 1982), host-parasite (Hawkins, 1992), or plant-pollinator. This partitioning is subjective, assuming that ecological and evolutionary dynamics among biotic interactions are independent from each other; this is currently evolving and has opened the way for theoretical frameworks to be developed and a more robust interpretation of network science. The study of ecological networks is traditionally based on the coupling between observed interaction networks, and subsequent

modeling and analysis of the structure. Visualizing networks as adjacency matrices is more common that force directed layouts, in which modules are easily visible.

Thanks to traditional ecological networks, several misleading conceptions about macro-organismal coevolution were squeezed out: (1) that coevolution leads to highly specific, directed and one to one interactions and that (2) coevolution within species-rich communities generates complicated assemblages that are intractable to generalization (Bascompte, 2010). Network analysis has shown that community interactions maximize robustness and functionality (Montoya et al., 2006; Thébault et Fontaine, 2010) such that interactions are fundamental units for understanding community dynamics, and productivity. Araújo et al., in 2011, suggested that species more exposed to climate change were poorly connected to other species of the network, while species more connected were less exposed. Ecological networks have helped defining keystone species, showing their disappearance could induce co-extinction and entire network collapses very rapidly, especially when confronted with climate change.

Mutualistic networks. They typically analyse plant-pollinator interactions, or plant-seed dispersers, by investigating how coevolutionary interactions shape species rich communities. Bascompte et al., 2007, summarised three features regarding mutualistic networks: (1) they are heterogeneous, meaning a majority of species interact with a few species, and a few species have a much higher number of interactions than what would be expected at random, (2) they are nested, in which specialist species interact with a subset of the group of species that generalists also interact with, and (3) they are asymmetric, so that a plant species will highly depend on an animal species for seed dispersal, but the animal weakly depends on the plant. It was proposed that asymmetries in these mutualistic coevolutionary networks would enhance long term coexistence and facilitate biodiversity maintenance (Bascompte et al., 2006). The effect of the interaction intimacy, whether high or low was studied, showing high intimacy mutualistic networks were nested whereas low intimacy mutualistic networks were modular (Guimaraes et al., 2007).

Antagonistic networks. Antagonistic networks often treat predator-prey networks, and host-parasitoid interactions. Food webs from different habitats and of different species richness have a scale-free network structure - where the degree distribution follows a power law with a parameter comprised between 2 and 3 - which makes them robust to disturbance such as extinction of species (Dunne et al., 2002; Kitano, 2002; Pascual et al., 2005). Intimacy in antagonistic networks also shapes the structure, between predators (low), parasites (intermediate) or parasitoid (high): from low to high, the network architecture changes from highly connected and weakly modular, to weakly connected and highly modular (Ings et al., 2009 for a review).

Merging different networks. Thébault and Fontaine, in 2011, opened the way to comparing different ecological networks based on models of population dynamics. Mutualistic plant-pollinator networks tend to be organized in nested patterns highly connected with increased stability. Antagonistic plant-herbivore networks tend to be stable when they are organized in compartments, and weakly connected. Antagonistic networks (studied through systems such as predator-prey or host-parasitoids) should be more specific than mutualistic ones, because the arms race between host and antagonist often leads to adaptation at the expense of the ability to attack alternative hosts, whereas in mutualistic interactions, organisms often specialize in traits shared by several species within a community, resulting in enhanced generalism. Later, Dunne (Dunne et al., 2013) investigated the reasons behind the altered structure of a food web after introduction of known parasites, asking if the perturbation was due to the true nature of parasites, or variation in diversity and complexity of the network. While most of the changes in the structure were related to scale-dependent phenomena (meaning that adding parasites or adding free living species is alike) changes in frequency of motifs of interaction among three taxa was parasite dependent.

Inferring biotic interactions from proxies. Morales-Castilla et al., 2015 proposed to construct biotic interaction networks based on proxies of macro-organisms, such as geographical data (species co-occurrence). In a multistep procedure, interacting groups of species are defined a priori based, for example, on trophic hierarchy (primary producers, grazers). This first step defined forbidden links, removing a large proportion of potential

links that could occur. This is clearly a hypothesis-driven approach, assuming our knowledge is sufficient to define those impossible interactions and that we know enough about trophic strategies regarding species. This is followed by calculation of the probabilities of interactions for the remaining links, through tricky estimation of strength and asymmetry of interactions.

To what extent can the analogy between ecological networks and microbial co-occurrence networks make sense? Can we really apply the macro-ecological network theory to the microbial world? The theoretical frameworks developed with terrestrial ecological network theory, and the questions they have asked, certainly resonate in the marine microbial world and our diatoms. Do phylogenetically related diatoms display similar interactions? How do diatom interactions impact diatom distribution? At what scales of space and time should diatom biotic interactions leave an imprint? How does the diatom network structure inform us on functions and stability of the ecosystem? Are the diatom species generalists or specialists in the predators or bacteria they co-occur with?

3.2. Diatoms act as repulsive segregators in the ocean

Target Journals for the following draft manuscript: Frontiers in Microbiology, Frontiers in Environmental Science, Environmental Microbiology.

Diatoms structure the plankton community based on repulsive segregation in the world's ocean

Flora VINCENT[1], Gipsi LIMA-MENDEZ[2], Karoline FAUST[2], Jeroen RAES[2], Chris BOWLER[1]

[1]Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR 8197, INSERM U1024, 46 rue d'Ulm, F-75005 Paris, France

[2]Department of Microbiology and Immunology, Rega Institute KU Leuven, Herestraat 49, 3000 Leuven, Belgium.

Abstract

Diatoms are a major component of phytoplankton, responsible for 25% of the annual primary production on Earth. As abundant and ubiquitous organisms, they are known to establish biotic interactions with many other members of the plankton realm, likely contributing to the global community structure and assemblages. Through the analysis of co-occurrence networks derived from the *Tara* Oceans expedition metagenomic data, and accounting for the importance of biotic and abiotic factors in shaping species' spatial distribution, we show that 13% of diatom pairwise associations are driven by environmental conditions. Diatoms act as repulsive segregators in the ocean, particularly towards harmful organisms such as potential predators and parasites. Genus level analysis supports the fact that abundant diatoms are not necessarily important in structuring the community, that exclusion patterns are species specific, and driven by characteristic distribution patterns. An extensive literature survey of diatom biotic interactions was compiled, of which 18.5% was recovered in the computed network. This is an exciting result that reveals the amount of what remains to be discovered in the field of planktonic biotic interactions.

3.2.1. Introduction

Marine microbes, composed of bacteria, archaea and protists, play an essential role in the functioning and regulation of the Earth's biogeochemical cycles (Falkowski et al., 2008). Their roles within planktonic ecosystems have typically been studied under the prism of bottom-up research, namely understanding how resources and abiotic factors affect their abundance, diversity and functions. On the other hand the effect of mortality, allelopathy, symbiosis and other biotic processes are also likely to shape their communities and to exert strong selective pressures on microbes (Strom, 2008). Indeed, with concentrations reaching $10^{\wedge 7}$/L protists and $10^{\wedge 9}$/L prokaryotes in a liter of seawater, biotic interactions are likely to impact community structure from the microscale to the ecosystem level.

Among marine protists, diatoms (Bacillariophyta) are of key ecological importance. They are a ubiquitous and predominant component of phytoplankton, characterized by their ornate silica cell walls, and responsible for approximately 40% of marine net primary productivity

(NPP; Nelson et al., 1995). The array of biotic interactions in which marine diatoms have been described is vast. They are fed upon by heterotrophic microzooplankton such as ciliates and phagotrophic dinoflagellates (Sherr and Sherr, 2007), as well as metazoan grazers such as copepods (Runge, 1988; Falkowski, 2002; Smetacek, 1998). Other well known interactions include symbiosis with nitrogen-fixing cyanobacteria (Foster et al., 2006), parasitism by chytrids and diplonemids (Gsell et al., 2013), diatom-targeted allelopathy by algicidal prokaryotes and dinoflagellates (Paul et al., 2011; Poulsen-Ellestad et al., 2014) and diatom-derived compounds detrimental to copepod growth (extensively reviewed in Pohnert, 2005). Nutrients also play a key role in the distribution and abundance of planktonic organisms. In the ocean, nutrients are often provided by upwelling processes that bring deep, cold and nutrient rich waters towards the surface. Diatoms typically thrive in high nutrient and high turbulent environments at the expense of the other major phytoplankton groups, dinoflagellates and haptophytes (Margalef, 1979). Competition for silica between diatoms and radiolarians, another silica utilizing member of the plankton has also been evoked (Harper et al., 1975).

Despite the strong biotic and abiotic selective pressures that likely influence diatom biogeography and evolution, they are considered as successful r-selected species (Armbrust, 2009). r-selection is an evolutionary strategy in which species can quickly produce many offspring in unstable environments, at the expense of individual “parental investment” and low probability of surviving to adulthood, such as rats. This is opposed to K-selection, in which species produce fewer descendants with increased parental investment such as elephants or whales (Pianka, 1970). Diatoms are one of the most diverse planktonic groups in terms of species, found to be widely distributed across the world’s sunlit ocean (Malviya et al., 2016) and capable of performing massive “blooms” in which diatom biomass can increase up to three orders of magnitude in just a few days (Platt et al., 2009). Their success has been attributed, in part, to a broad range of predation avoidance mechanisms (Irigoien, 2005) such as their solid mineral skeleton (Hamm et al., 2007), chain and spine formation in some species, and toxic aldehyde production. However, a global view of their capacity to interact with other organisms and an assessment of mechanisms shaping contemporary community structures are still lacking.

Co-occurrence networks using meta-omics data are increasingly being used to study microbial communities and interactions (Faust and Raes, 2012; Li et al., 2016), e.g., in human and soil microbiomes (Faust et al., 2012; Barberan et al., 2012) as well as in marine and lake bacterioplankton (Fuhrman et al., 2008; Eiler et al., 2011; Milici et al., 2016). Such networks provide an opportunity to extend community analysis beyond alpha and beta diversity towards an understanding of the relational roles played by different organisms, many of whom are uncultured and uncharacterized (Proulx et al., 2005; Chaffron et al., 2010). Over large spatial scales, non random patterns according to which organisms frequently or never occur in the same samples are the result of several processes such as biotic interactions, habitat filtering, historical effects as well as neutral processes (Fuhrman, 2009). Quantifying the relative importance of each component is still in its infancy. However, these networks can be used to reveal niche spaces, to identify potential biotic interactions and to guide more focused studies. Much like in protein-protein networks, interpreting microbial association networks also relies on literature-curated gold standard databases (Li et al., 2016), although such references do not yet exist for most planktonic groups.

As part of the recent *Tara* Oceans expedition (Karsenti et al., 2011; Bork et al., 2015), determinants of community structure in global ocean plankton communities were assessed using co-occurrence networks (Lima-Mendez et al., 2015). Pairwise links between species were computed based on how frequently they were found to co-occur in similar samples (positive correlations), or on the contrary if the presence of one organism negatively correlated with the presence of another (negative correlations). In order to prevent spurious correlations due to the presence of a third confounding factor such as abiotic factors, interaction information was furthermore calculated to assess whether or not edges were driven by an environmental parameter. The *Tara* Oceans Interactome represents an ideal case to investigate global scale processes involving diatoms, as it maximizes spatio-temporal variance across a global sampling campaign and captures system-level properties.

Here, we reveal diatoms to be the organismal group with proportionally the most negative interactions within the *Tara* Oceans Interactome. We then hypothesize that as good competitors, diatoms are likely to act as repulsive segregators – they display more negative than positive associations (Morueta-Holme et al., 2016) - in particular against potentially harmful organisms in the global ocean, and ask what are the underlying distribution patterns that drive these interactions. By extracting knowledge from a recent biogeography study of diatoms (Malviya et al., 2016), we show that abundant diatoms are not necessarily key players in the community structure. We then performed an extensive literature survey of current knowledge regarding diatom biotic interactions, revealing that it is highly skewed towards predation interactions and macroorganisms. Finally, by combining empirical knowledge and data-driven studies of microbial associations, we reveal important yet understudied players of the diatom environmental interactome.

3.2.2. Results

Diatoms are repulsive segregators in the open ocean.

Co-occurrence amongst different micro-organisms in the plankton has recently been investigated at a large scale by Lima-Mendez et al., 2015 (Annex D). The resulting interactome represents species (nodes), connected by links (edges) that represent either positive or negative associations. Positive associations should be understood as two organisms that are often abundant in the same sample and negative associations as two organisms that are rarely abundant in the same sample. Due to the potential impact of environmental drivers on the co-occurrence of two organisms, the *Tara* Oceans Interactome also provides insight into the effects of abiotic factors on pairwise correlations. This assessment provides the opportunity to disentangle biotic from abiotic factors at the pairwise level.

The global ocean interactome reports over 90,000 statistically significant correlations, with ~68,000 of them being positive, ~ 26,000 of them being negative, and ~ 9,000 due to the simultaneous higher correlation of two organisms (OTUs) with a third environmental parameter. Diatoms are involved in 4,369 interactions, making them the 7th most

connected taxonomic groups after syndiniales (MALVs), arthropods, dinophyceae, polycystines, MASTs and prymnesiophyceae. All groups except diatoms and polycystines display a higher number of positive edges than negative ones **(Table S1)**. Overall, diatoms represent around 3% of all the positive associations (2,120/68,856) and 9.5% of all negative associations (2,249/23,777) showing that their contribution to negative associations is much higher than their contribution to positive co-occurrences, unlike major taxonomic groups involved in the interactome. The positive to negative ratio provides a measure of the species role in the network, and suggests that diatoms act as repulsive segregators, meaning they have more negative than positive associations (Morueta-Holme et al., 2016). Amongst all the pairwise associations involving diatoms and other organisms in the plankton (N=4,369), only 13% were due to a third environmental parameter illustrating a shared preference for a particular abiotic condition (N=566), leaving 87% of the associations solely explained by the abundance of the two organisms **(Figure S3.3)**.

A finer analysis revealed the major taxonomic groups with which diatoms correlate or anti-correlate. Positive correlations involve mainly arthropoda (9.2% of diatom negative correlations), dinophyceae (8.7%), and syndiniales (11.7%). Negative correlations include the three previous groups – arthropoda (11.5%), dinophyceae (11.3%), syndiniales (11.1%) – as well as the polycystina (6%), a major group of radiolarians that produce mineral skeletons made from silica. We investigated whether or not these patterns were consistent across the ocean, asking if the other taxonomic groups of plankton displayed similar patterns with these organisms. The number of negative correlations involving diatoms with copepods, syndiniales, dinoflagellates and radiolarians was much higher than what would be expected at random based on binomial testing **(Table S2, Figure S3.4-S3.6)**. For example, out of the 64 groups that display at least 10 associations with copepods, only 6 anti-correlate more than they co-occur with metazoan predators **(Figure 3.7)**.

Sub-networks were then extracted for both positive and negative associations involving diatoms with syndiniales (MALVs), MASTs, dinophyceae, and copepods **(Figure 3.8)**. The clustering coefficient measures the tendency of a graph to form clusters compared to a random network, and it was higher in positive association networks than in negative

association networks, meaning that the positive co-occurrence networks tend here to form more modules than the negative ones. In network biology, two nodes are connected if there is a path of edges between them, so all nodes that are pairwise connected form a connected component. Consequently, low connected components relate to strong connectivity because many nodes are connected. Connected components were higher in MALV and MAST sub-networks, and lower in dinophyceae, meaning that the former sub-networks had lower connectivity. Associations with MASTs were more centralized, meaning that the sub-network has a star-like topology in which nodes of the network, on average, do not have the same connectivity and are not uniformly connected **(Figure S3.7).** In addition, many MAST nodes belonged to the MAST-3 clade, known to harbor the diatom parasite *Solenicola setigara* (Gómez et al., 2011). In order to investigate the strength of repulsive interactions, average Spearman correlations of the values were computed **(Table S3).** Contrary to expectations, they were higher for copresences than for exclusions. However, we compared diatom Spearman scores with those of polycystines as control group and found that diatoms have stronger negative scores, reflecting a higher potential as repulsive segregators with respect to potential harmful organisms. Average scores were higher for MASTs (-0.66+-0.09) and MALVs (-0.59+-0.09).

Global-scale genus abundance does not determine importance in community structuring

While abundant diatoms are likely to be important players in biogeochemical cycles such as NPP (Net Primary Production) and carbon export, are they also relevant in structuring plankton communities? To address this question, the ten most abundant diatom genera (based on 18S rDNA V9 read abundance) were analyzed with respect to their positions in the interactome **(Table S4)**. This analysis revealed that no significant correlation was found between the total abundance of the genus, and the number of edges the genus is involved in (Spearman p.value = 0.96), nor the number of nodes involved (Spearman p.value = 0.45). Some barely play a role in the interactome, for example *Attheya* is the tenth most abundant genus (96,926 reads), yet is only represented in 10 edges across the interactome. On the other hand, the diatom genus *Synedra,* that is not notably abundant at the global level (ranked as the 22nd most abundant diatom with 28,700 reads), was involved in over 100

significant associations. *Leptocylindrus* (710,295 reads), is involved in 667 interactions **(Figure S3.8)**.

Statistics of network level properties provide further insights into the overall structure of species-specific assemblages, and was investigated at the genus level for *Leptocylindrus*, *Proboscia*, and *Pseudo-Nitzschia*, each of which displayed high clustering coefficients meaning neighbors of nodes are connected between each other. Similarly, these genera displayed a higher average number of neighbors, meaning that the average connectivity of the nodes in the network was higher, which could be interpreted as species being generalist in their interactions. On the other hand, the *Chaetoceros*, *Eucampia*, and *Thalassiosira* sub-networks displayed larger diameters, thus inferring the largest distance between two nodes. This is illustrated by the fact that a few diatom nodes are connected both positively and negatively to a large number of nodes that are not connected to any diatom partners, therefore a more specific type of behavior with respect to interactions. **(Figure S3.9)**.

Species level repulsive segregation determined by blooming strategies

Due to the small number of individual barcodes involved at the species level, we decided to conduct a finer analysis and ask whether or not different barcodes of the same (abundant) genera displayed specificity in the type of interactions and partners they interact with. We illustrate this barcode specificity with three different genera: *Chaetoceros*, *Pseudo-nitzschia* and *Thalassiosira*. *Chaetoceros* interactions reveal that different species display very different co-occurrence patterns **(Figure 3.9).** The barcode ”29f84ed97c31eabfc3e787fd686442a4,” assigned to *Chaetoceros rostratus* is essentially only involved in positive co-occurrences, while barcode “8fd6d889852840ef8c8863cebdc14d10,” assigned to *Chaetoceros debilis,* is the major driver of negative associations involving dinophyceae, MASTs, syndiniales and arthropods. This could reflect the different species tolerance to other organisms, since several *Chaetoceros* species are known to be harmful to aquaculture industries (Albright et al., 1993); *Chaetoceros debilis* in particular can cause physical damage in fish gills (Kraberg et al., 2010).

Pseudo-nitzschia barcodes are primarily involved in positive correlations. However, they display exclusions with organisms such as arthropoda and dinophyceae, and are known to

produce toxic domoic acid in the presence of copepods (Tammilehto et al., 2015). No exclusions regarding syndiniales appear, although barcode-level specificity is observed with "1d16c182297bf5aa52f3d99c522ade94," which is involved in a much higher number of interactions than "b56c311cb937f906f5fd96ee5983fcac." Unfortunately, diatoms were not assigned at the species level. Finally, the *Thalassiosira* sub-network displays major negative associations with syndiniales, arthropoda, and polycystines, with one of the three representative barcodes being highly involved in structuring the community (53bb764df052b8ba43d74f8b4e3b7e92). The abundance of the barcodes aforementioned, involved in a high number of mutual exclusions, is typical to that of blooming diatoms (high relative abundance in one specific sample). This observation was confirmed by analyzing the distribution patterns of top diatom barcodes involved in exclusions such as 90dade88b591756bdc889e60c5c6d424 (Unassigned *Bacillariophyta* blooming in Indian Ocean Station 36), 4c4a832b7b9a29675ef3bd9bc394adbf (*Raphid-Pennate*, Marquesas Station 122), 8fd6d889852840ef8c8863cebdc14d10 (*Chaetoceros* in Southern Ocean Station 88), 30191c0570b3035d38a5bc7cc8738a04 (*Actinocyclus* in Indian Ocean Station 33), ae808698d6131569f7b5a8abd09f497a (*Proboscia*, Station 116), 53bb764df052b8ba43d74f8b4e3b7e92 (*Thalassiosira* in Indian Ocean Station 36).

Segregation towards heterotrophic prokaryotes is not significant

We hypothesized that repulsive segregation by diatoms could also apply to heterotrophic bacteria, leading to higher than at random negative associations towards heterotrophic than autotrophic bacteria. Diatom - prokaryote associations represent 19% of the whole diatom co-occurrence network which is considered as average when compared to Bacteria associations in copepod interactions (28%), dinophyceae (18.5%), radiolarian (20.5%) and syndiniales (16.3%). Bacteria were classified according to their primary nutritional group based on the literature. No significant exclusion toward heterotrophic bacteria was found based on the current data **(Figure S3.10)**.

A skewed knowledge about diatom biotic associations

To review current knowledge about diatom interactions we generated a database that assembled the minable knowledge in the literature about diatom associations from both

marine and freshwater habitats. A total of 1,533 associations from over 500 analyzed papers involving 83 unique genera of diatoms and 588 unique genera of other partners are reported here, illustrating the diversity of association types such as predation, symbiosis, allelopathy, parasitism, and epibiosis, as well as the diversity of partners involved in the associations, including both prokaryotes and eukaryotes, micro- and macro-organisms **(Figure 3.6)**.

We noted that 58% (883 out of 1,533) of the interactions are labelled "eatenBy" and involve mainly insects (267 interactions; 30% of diatom predators) and crustaceans (15% of diatom predators). Cases of epibiosis, representing approximately 10% of the literature database, were largely dominated by epiphytic diatoms living on plants (40% of epibionts) and epizoic diatoms living on copepods (9% of epibionts). Parasitic and photosymbiotic interactions, although known to have significant ecological implications at the individual host level as well as at the community composition scale (Veen, 2008), represented only 15% of the literature database for a total of 219 interactions, involving principally diatom associations with radiolarians and cyanobacteria. Interactions involving bacteria represent 72 associations (4.8 % of the literature database).

The distribution of habitats amongst the studied diatoms reveals a singular pattern: the majority of diatom interactions in the literature are represented by a handful of freshwater diatoms, whereas many marine species are reported in just a small number of interactions **(Figure S3.1)**. In terms of partners involved **(Figure S3.2)**, one third are represented by insects feeding upon diatoms in streams and crustaceans feeding upon diatoms in both marine and freshwater environments. Other principal partners are plants, upon which diatoms attach as "epiphytes," such as Posidonia (seagrass), Potamogeton (pondweed), Ruppia (ditchgrass) and Thalassia (seagrass). Consequently, our knowledge based on the literature produces a highly centralized network containing a few diatoms mainly subject to grazing. Major diatom genera for which interactions are reported in the literature are *Chaetoceros spp* (215 interactions, marine and freshwater), *Epithemia sorex* (135 interactions, freshwater), and *Cymbella aspera* (115 interactions, freshwater).

Overlapping empirical evidence on data driven results reveals gaps in knowledge and extends it to the global ocean

In order to go towards edge annotation of the co-occurrence network, the literature database presented here was used. The occurrence of a specific genera in the literature was compared to its occurrence in the *Tara* Oceans interactome. On average, the co-occurrence network revealed much more potential links between species than what was reported in the literature. Disparity was especially high for *Pseudo-Nitzschia*, mentioned in 17 interactions in the literature compared to 307 associations in the interactome. On the other hand, many diatoms involved in several associations in the interactome are absent from the literature, such as *Proboscia* and *Haslea* **(Figure S3.11& S3.12)**.

Out of 1,533 literature-based interactions, 178 could potentially be found in the *Tara* Oceans Interactome, as both partners had a representative barcode in the *Tara* Oceans database. A total of 33 literature-based interactions (18.5% of the literature associations) were recovered in the network at the genus level, representing a total of 289 interactions from the interactome and 209 different barcodes. These 289 interactions represent 6.5% of all the associations involving Bacillariophyta in the *Tara* Oceans co-occurrence network. By mapping available literature on the co-occurrence network, we can see that the major interactions recovered are those involving competition, predation and symbiosis with arthropods, dinoflagellates and bacteria **(Figure 3.10)**. However predation by polychaetes and parasitism by cercozoa and chytrids are missing from the interactome.

3.2.3. Discussion

We provide a detailed study of the diatom interactome based on co-occurrence networks, derived from the largest spatial metagenomic survey conducted to our knowledge, combined with an up-to-date review of current empirical knowledge regarding diatom biotic interactions. Our literature survey reveals a skewed knowledge, focusing on freshwater diatoms and predation by macro-organisms, with very few parasitic, photosymbiotic or bacterial associations. The relative poverty of marine microbial studies can be explained by

the difficulty of accessing these interactions in the field, which obviously limits our understanding of how such interactions structure the community at global scale.

Out of the complete diatom association network extracted from the *Tara* Oceans Interactome, co-localization and co-exclusion of diatoms with other organisms are due to shared preferences for an environmental niche in 13% of the cases, emphasizing the importance of biotic factors in 87% of the associations. By excluding major functional groups such as predators, parasites and competitors, diatoms act, at a global scale, as a repulsive segregator preventing their co-occurrence in similar samples (Smetacek, 2012). Diatoms are known to have developed an effective arsenal composed of silicified cell walls, spines, toxic oxylipins, and chain formation to increase size, so we propose that the observed exclusion pattern reflects the worldwide impact of the diatoms arms race against harmful organisms. Additionally, building upon the phylogenetic affiliation of individual sequences, barcodes can be assigned to a plankton functional type that refers to traits such as the trophic strategy and role in biogeochemical cycles (Quéré et al., 2005). As demonstrated in the *Tara* Oceans interactome (Lima-Mendez et al., 2015), diatoms compose the "phytoplankton silicifiers" metanode, and display a variety of mutual exclusions that distinguish them from other phytoplankton groups.

Sub-network topologies reveal that MASTs - diatom and MALV - diatom networks all display lower connectivity, suggesting more specialist interactions than with copepods or dinophyceae. MAST networks were also more centralized, showing non-uniform connections among components and higher specificity in the diatom - marine stramenopile links. Correlation values are often neglected in co-occurrence analysis but here they reveal stronger exclusion patterns of diatoms against MASTs and MALVs. Exclusion between diatoms and MASTs is therefore more specific, and stronger, than compared to copepods or dinoflagellates. Analysis at the genus level shows that abundant diatoms such as *Attheya* do not play a central role in structuring the community, contrary to *Synedra* that, at a global scale, had small significance in terms of abundance but is highly connected to the plankton community. We show the existence of a species level segregation effect, that can be attributed to harmful traits (Kraberg et al., 2010), reflected by blooming distribution

patterns for the top repulsive segregating diatoms. If diatom blooms are known to be triggered by light and nutrient perturbation, the negative associations were not driven by a third environmental parameter emphasizing the biotic component of the segregative effect.

Comparing empirical knowledge and data-driven association networks reveal understudied genera such as *Leptocylindrus*, or *Actinocyclus*, and those that are not even present in the literature such as *Proboscia* and *Haslea*. However, we stress that *Proboscia* is a homotypic synonym of *Rhizosolenia* that is found in the interactome, which gives rise to the long lasting debate about non-universal taxonomic denomination and its incidence on diversity analysis.

18.5% of the literature database was recovered in the interactome, however this explained only 6.5% of the 4,369 edges composing the diatom network. The gap between the 20% diatom-bacteria interactions in the *Tara* Oceans Interactome and 4.8% diatom-bacteria associations described in the literature highlights how little we know about host-associated microbiomes at this stage. In many ways, this high proportion of unmatched interactions should be regarded as the "Unknown" proportion of microbial diversity emerging from metabarcoding surveys. Part of it is truly unknown and new, part of it is due to biases in data gathering and processing and part of it is due to the lack of an extensive reference database.

Many challenges remain regarding the computation, analysis, and interpretation of co-occurrence networks despite their potential to uncover major processes shaping diatom-related microbial communities. Recent studies are exploring the methodological bias of each co-occurrence method by attempting to detect specific associations within mock communities (Weiss et al., 2016) whilst others admit that applying network statistics to microbial relationships is submitted to high variability depending on taxonomic level of the study and criteria used to compute networks (Williams et al., 2014). Assigning biological interactions such as predation, parasitism or symbiosis to correlations is still cumbersome and will require both proper references of biotic interactions (Li et al., 2016), and further studies that investigate dynamics of interactions through space and time, and their sensitivity to co-occurrence detection. Attempts to develop a theoretical framework for species interaction detection based on co-occurrence networks is underway (Morales-

Castilla et al., 2015; Cazelles et al., 2015; Morueta-Holme et al., 2015), but must be confronted with actual *in situ* data. Furthermore, a vast body of literature already exists in the field of ecological networks, traditionally focusing on observational non-inferred data and the modeling of foodwebs, host-parasite and plant-pollinator networks (Ings et al., 2009; Bascompte et al., 2010). Various properties linked to the architecture of these antagonistic and mutualistic networks have been formalized, such as nestedness, modularity, or the impact of combining several types of interactions in a single framework (Thébault et al., 2010; Fontaine et al., 2011). As the fields of ecological networks and co-occurrence networks both focus on biotic interactions and how they structure the community, more cross fertilization between the two disciplines would highly benefit both communities, ultimately helping to understand the laws governing the "tangled bank" (Darwin, 1859).

3.2.4. Materials and methods

Construction of Diatom Interaction Literature Data Base. Literature was screened to look for all ecological interactions involving diatoms to establish the current state of knowledge regarding the diatom interactome, both in marine and freshwater environments. Diatom ecological interactions as defined in this paper are a very large group of associations, characterized by (i) the nature of the association defined by the ecological interaction or the mechanism (predation, symbiosis, mutualism, competition, epibiosis), (ii) the diatom involved, and (iii) the partners of the interaction.

The protocol to build the list of literature-based interactions was the following (i) collect publications involving diatom associations using (a) the Web of Science query TITLE : (diatom*) AND TOPIC : (symbio* OR competition OR parasit* OR predat* OR epiphyte OR allelopathy OR epibiont OR mutualism) ; (b) Eutils tools to mine Pubmed and extract ID of all publications with the search url http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=diatom+symbiosis&usehistory=y and the same keywords; (c) the get_interactions_by_taxa(sourcetaxon = "Bacillariophyta") function from the RGlobi package (Polen et al., 2014), the most recent and extensive automated database of biotic interactions; and (d) personal mining from other publication browsers and input from experts (ii) extract when relevant the partners of

the interactions based on the title and on the abstract for Web of Science, Pubmed and personal references and normalize the label of the interaction based on Globi nomenclature (iii) display KRONA plot with Type of Interaction / Partner Class / Diatom genus / Partner genus_species as shown in **Figure 3.6**. Cases of episammic (sand) and epipelon (mud) interactions were not considered as they involved association with non-living surfaces.

Relative proportion of co-occurrences and exclusions with respect to major partners and network analysis. All analyses were performed on the published co-occurrence network in Lima-Mendez et al., 2015. Environmental drivers of diatom related edges are available in **Table S5**. Four independent matrices were created from the interactome regarding the major partners interacting with diatoms (copepods, dinophyceae, syndiniales and radiolaria) containing only pairwise interactions that involved the major partner and binomial testing was done using the dbinom and pbinom function as implemented in the {stats} package of R version 3.3.0. Subnetwork topologies were analyzed using the NetworkAnalyzer plugin in Cytoscape (Shannon et al., 2003) as described in (Doncheva et al., 2012). Network topologies for major groups are available in **Table S6**.

Major diatom interactions. The 10 most abundant diatom genera in the surface ocean were selected based on the work published by (Malviya et al., 2016). Network topologies are available in **Table S7**. Their co-occurrence network was then extracted from the global interactome and analyzed at the ribotype level. Distribution of individual barcodes was assessed across the *Tara* Oceans sampling stations.

Comparison of literature interactions and diatom interactome. All partner genera interacting with diatoms based on the literature were searched for in the *Tara* Oceans dataset based on the lineage of the barcode. For each barcode that had a match, identifiers (“md5sum”) were extracted creating a list of 954110 barcodes to be searched for in the global interactome.

3.2.5. Figures and supplementary material

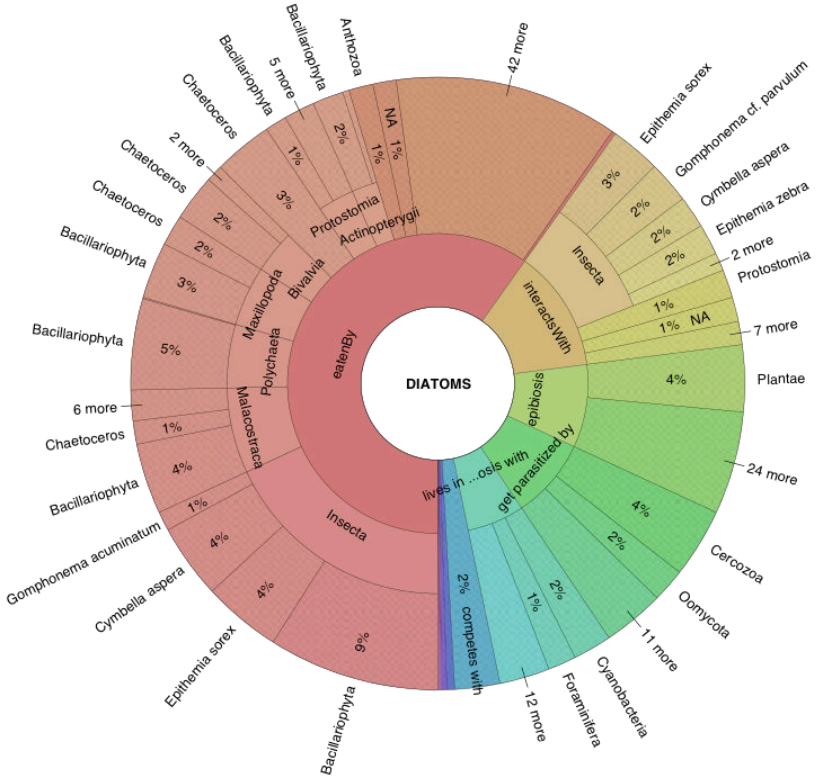


Figure 3.6. Current knowledge of diatom biotic interactions.
KRONA plot based on available literature concerning diatom associations, mined and manually curated from Web Of Science, PubMed and Globi (**Table S8**).

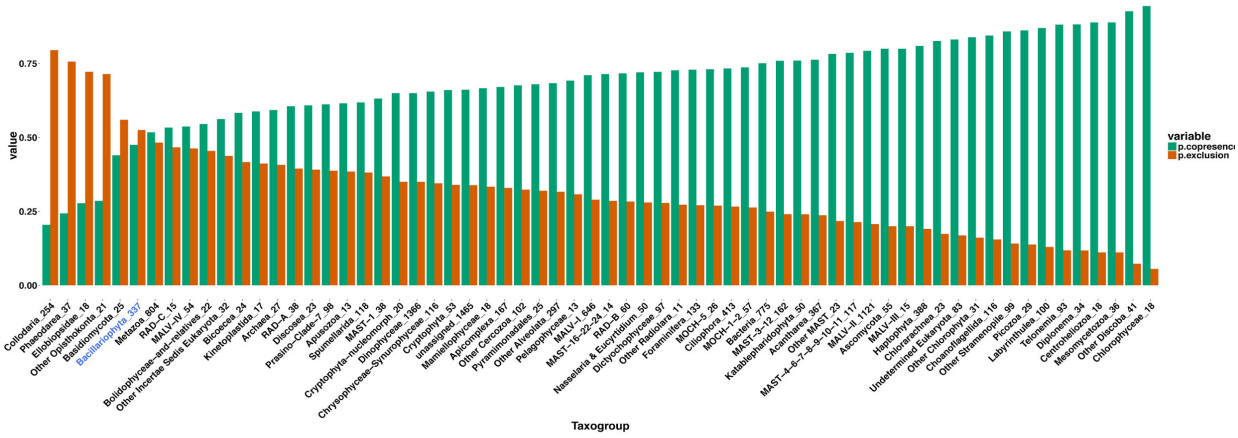


Figure 3.7. Relative proportion of exclusions and co-occurrences of copepods.
All interactions involving copepods were extracted from the *Tara* Oceans interactome, and only taxonomic groups with over 10 edges were kept. For each group, the relative proportion of positive (green) and negative (red) edges are displayed, and absolute number of edges involved is indicated in the abscissa. Diatoms (Bacillariophyta) are highlighted in blue.

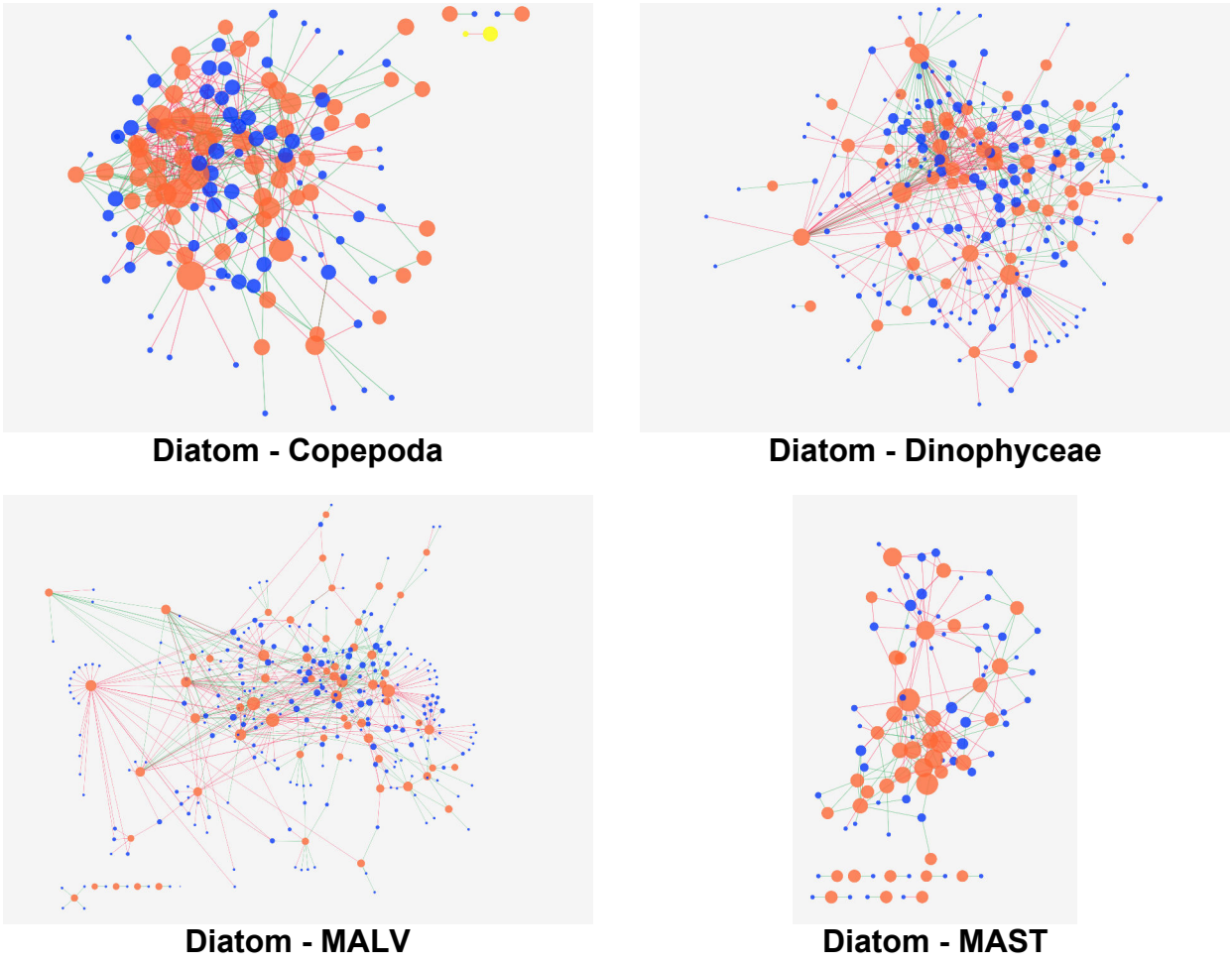


Figure 3.8. Subnetwork topology of diatoms and major partners.
Diatom nodes are colored in orange, and the corresponding partner nodes are colored in blue. Green edges correspond to positive co-occurrences while red edges correspond to negative correlations. The size of the node corresponds to a continuous mapping of the degree in the global diatom interactome. Corresponding subnetwork descriptors are available in **Table S6**

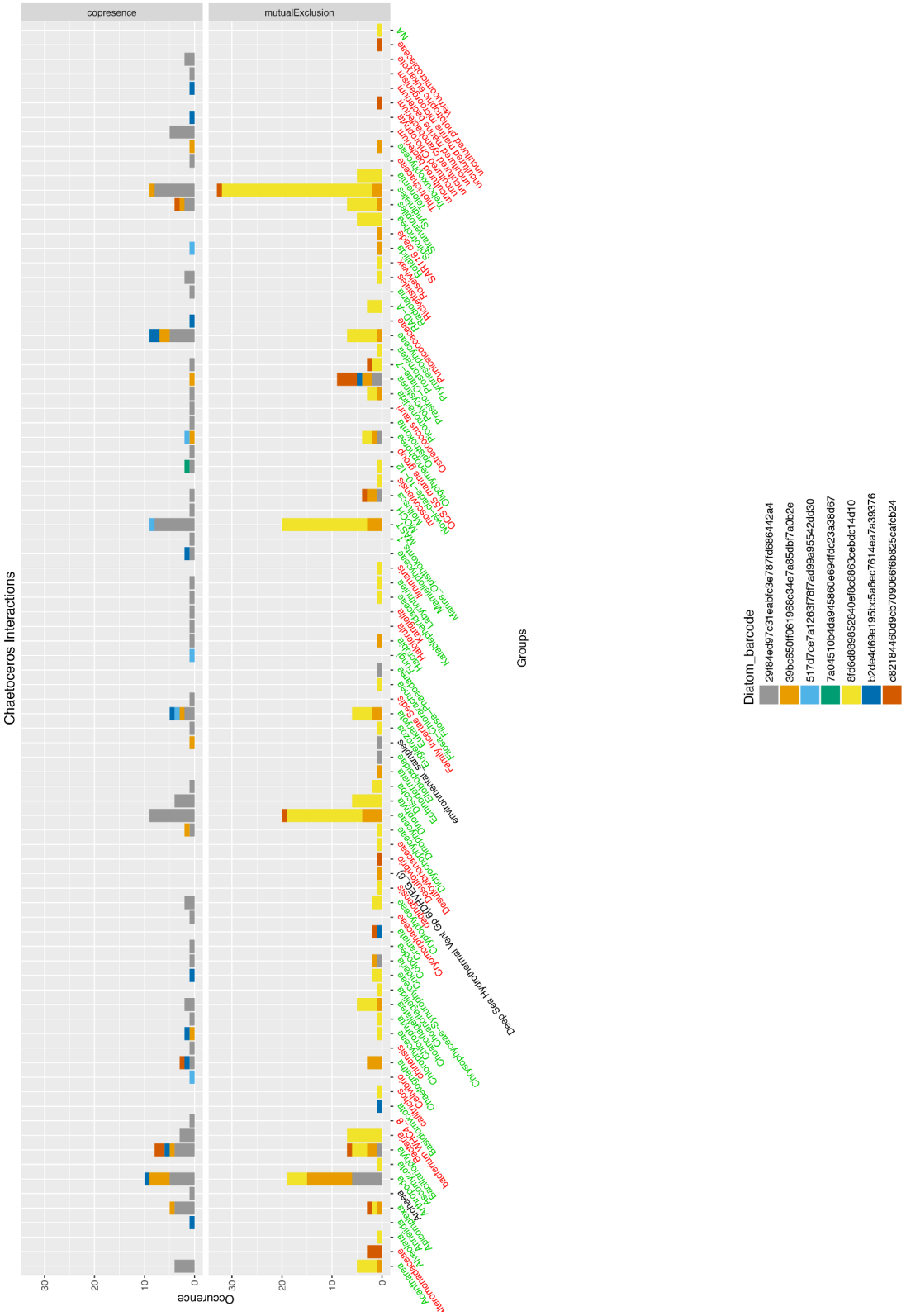


Figure 3.9. Barcode level associations of the diatom genus Chaetoceros.
Interactions are colored by the corresponding barcodes that are named according to the identifier in the *Tara* Oceans metabarcoding data. Partners of interactions are coloured by domain of life (Green: Eukaryotes, Red: Prokaryotes, Black: Archaea). The higher pannel represents positive associations, the lower panel represents exclusions.

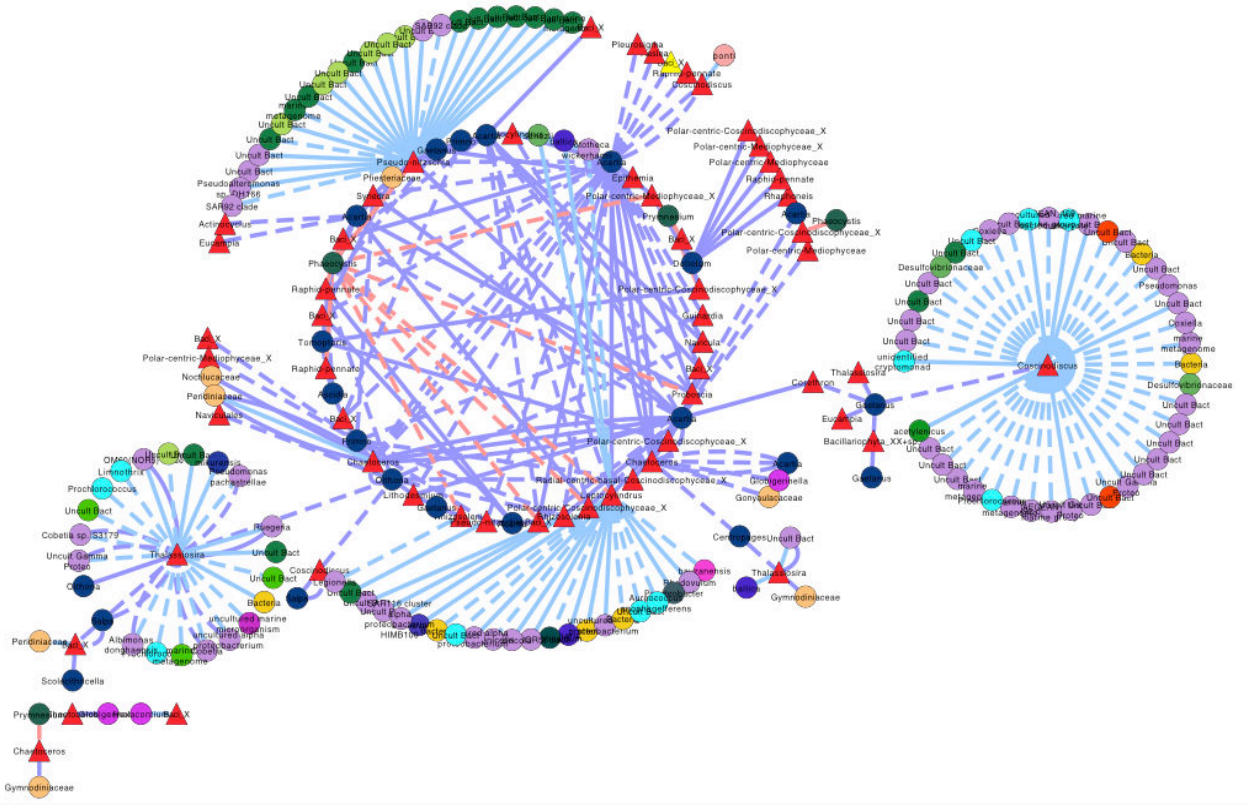


Figure 3.10. Literature confirmed associations from the *Tara* Oceans interactome.
The biotic interactions known from the literature were searched from in the *Tara* Oceans interactions. Dotted lines represent negative correlations in the interactome, full lines represent positive correlation in the interactome. Edges are colored by their respective taxonomic labelling in the literature (Purple: predation; Blue: Symbiosis/Parasitism; Red: Competition). Diatoms are represented in red triangles (**Table S9 for the whole list of recovered edges**).

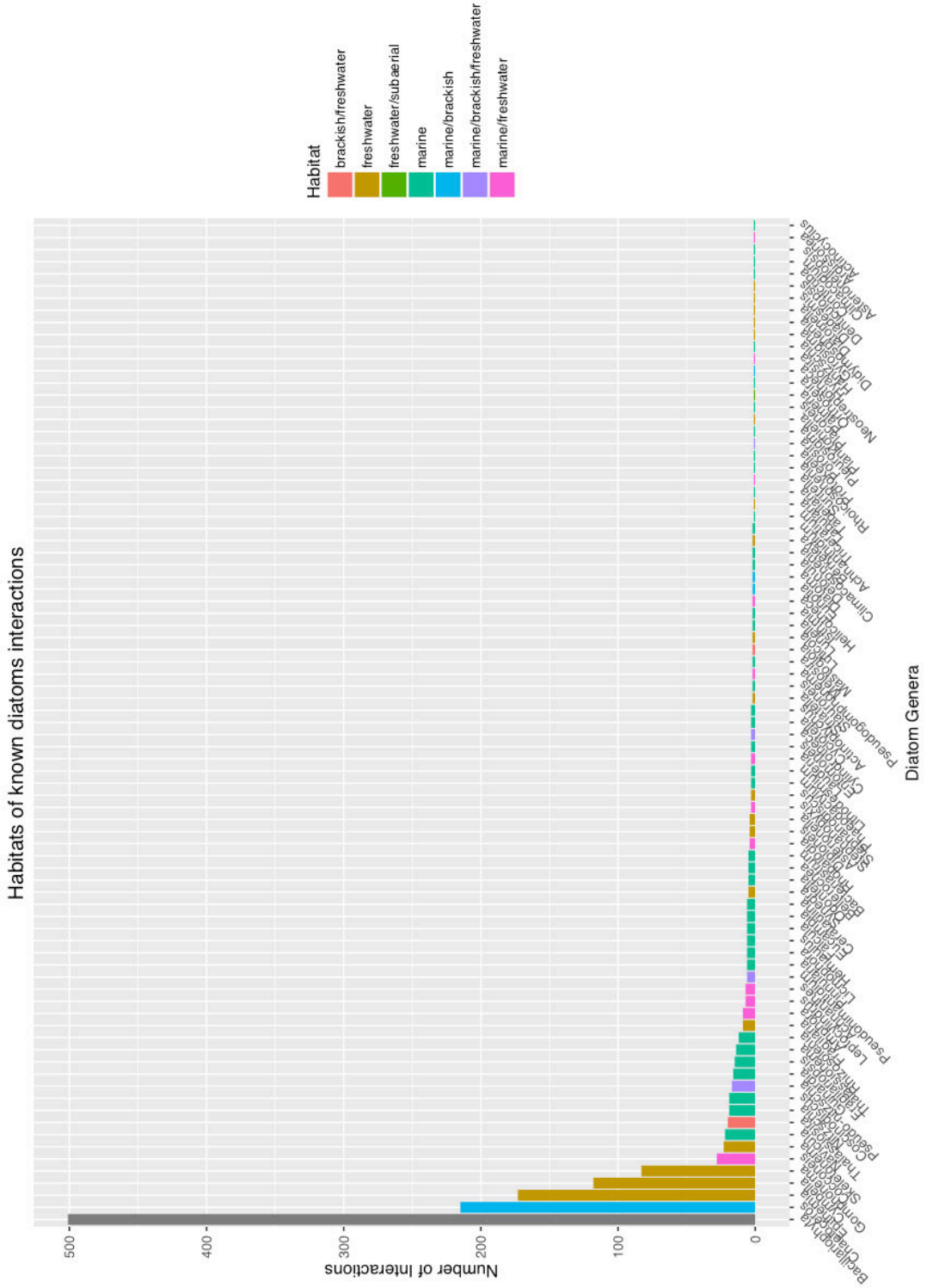


Figure S3.1. Habitats of diatoms involved in known interactions.
For each available diatom genera in the literature interaction database, habitat was assigned based on available knowledge.

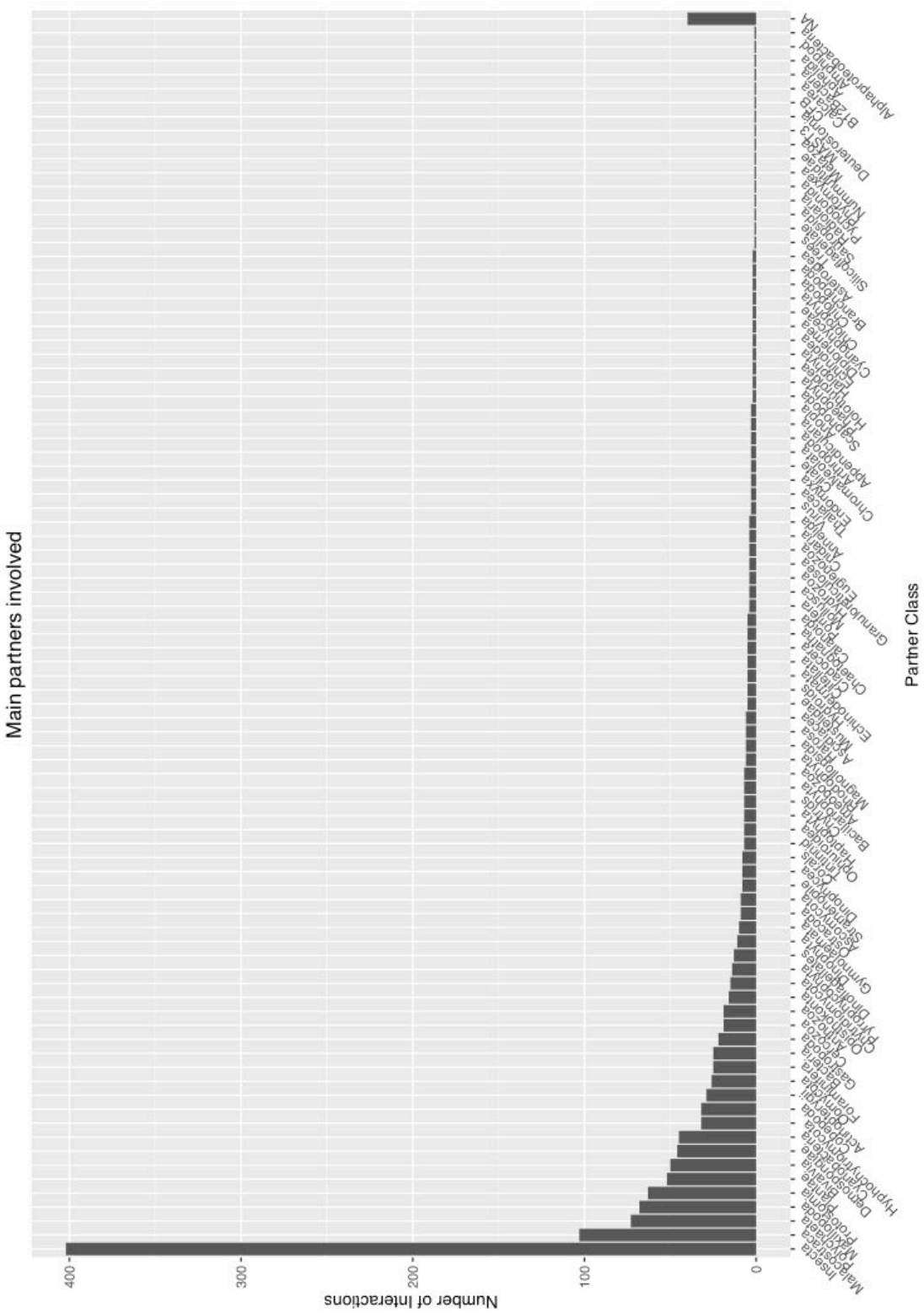


Figure S3.2. Main partners involved in diatom interactions based on the literature.

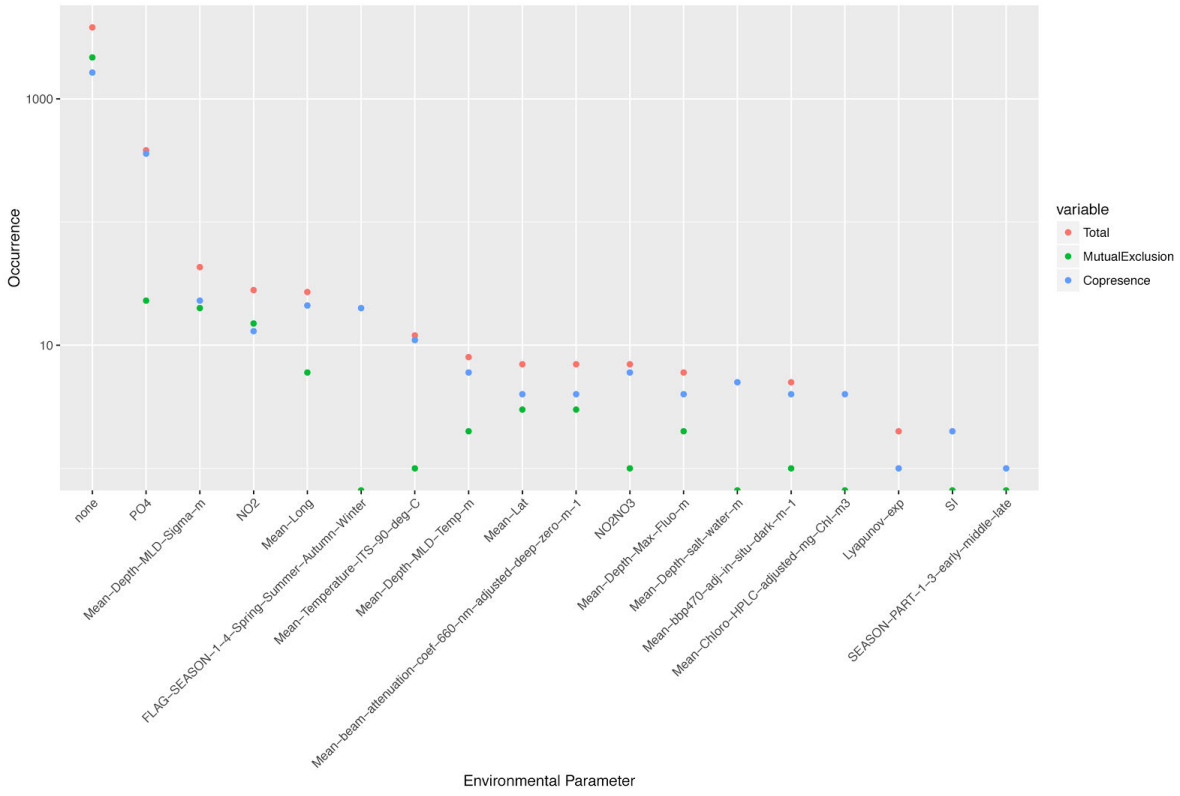


Figure S3.3. Major environmental drivers of diatom edges.
The methodology used to disentangle the relative influence of abiotic factors on diatom interactions is available in Annex D under "Methods/Indirect taxon edge detection" and was implemented by K.Faust. Environmental Parameters: Phosphate (PO4); Mixed Layer Depth (MLD, layer in which active turbulence homogenizes water, estimated by density – sigma- and temperature); Nitrite (NO2); Light scattering by suspended particles (Beam attenuation 660nm); Backscattering coefficient of particle (bbp470); HPLC Chlorophyll pigment measurement (HPLC-adjusted); Ocean perturbation (Lyapunov exponent), Silicate (SI) and categorical variable for season. A full description of the environmental parameters is available on the PANGEAE website (https://doi.pangaea.de/10.1594/PANGAEA.840718).

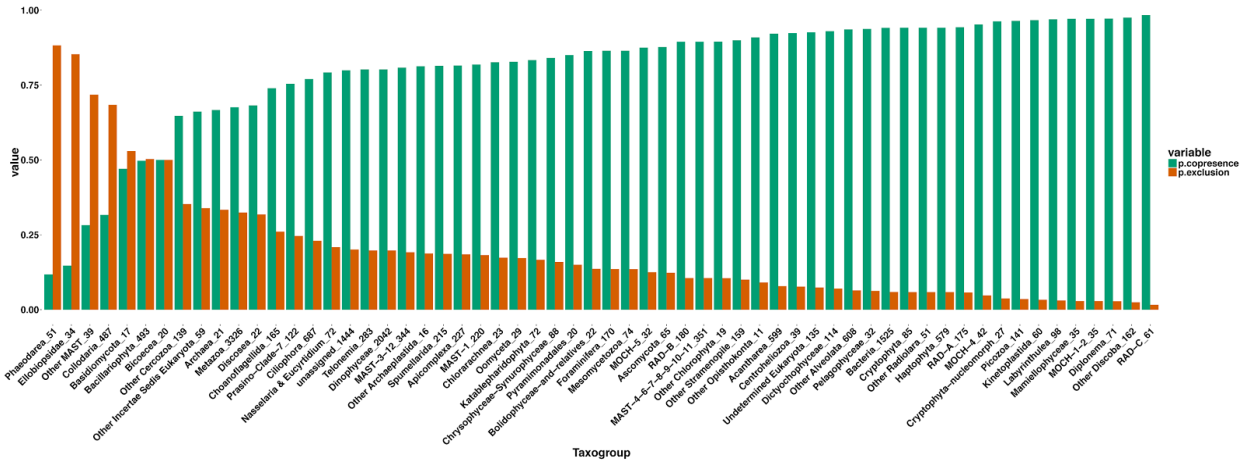


Figure S3.4. Relative proportion of positive and negative interactions for syndiniales.
All the interactions involving syndiniales were extracted from the *Tara* Oceans Interactome (Lima-Mendez, 2015) and the relative proportion of positive (green) and negative (red) interactions was computed for each major taxonomic group and plotted by decreasing negative edges. The absolute number of interactions between the taxonomic group and syndiniales is given in the legend. Diatoms are the 6th group from the left. Refer to **Table S2**

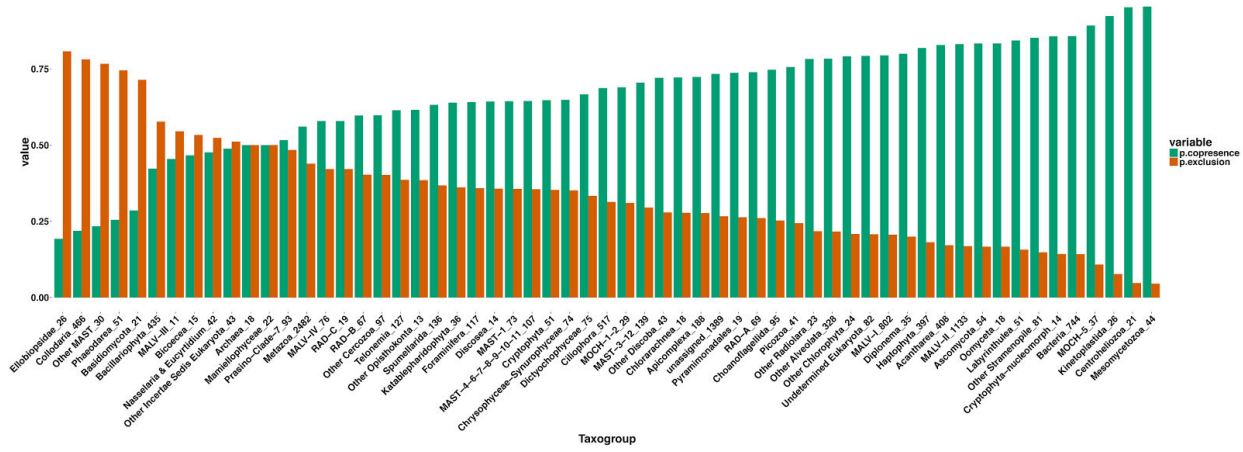


Figure S3.5. Relative proportion of positive and negative interactions for dinophyceae.
All the interactions involving dinophyceaes were extracted from the *Tara* Oceans Interactome (Lima-Mendez et al., 2015) and the relative proportion of positive (green) and negative (red) interactions was computed for each major taxonomic group and plotted by decreasing negative edges. The absolute number of interactions between the taxonomic group and Dinophyceaes is given in the legend. Diatom are the 6th group from the left.

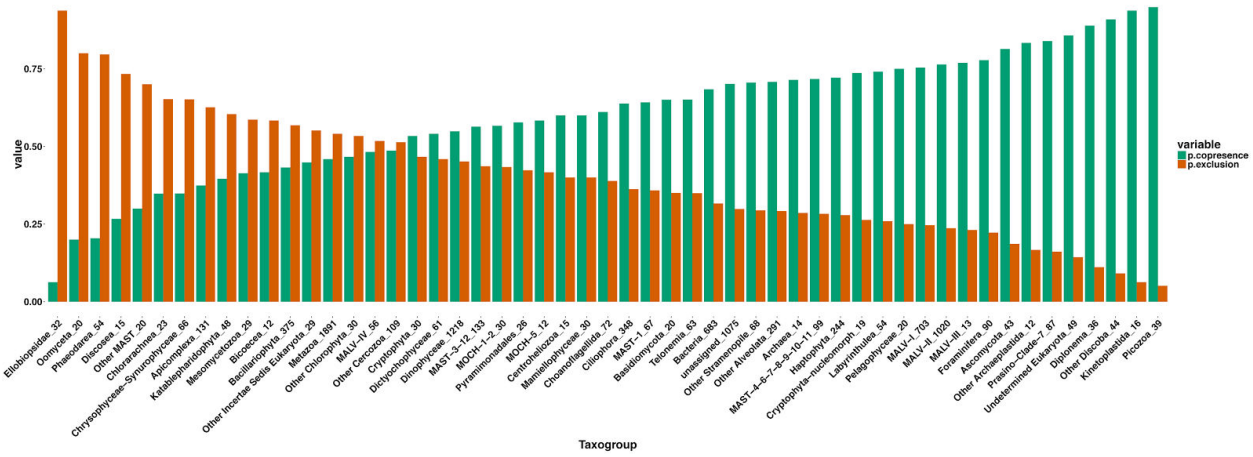


Figure S3.6. Relative proportion of positive and negative interactions for radiolaria.
All the interactions involving radiolaria were extracted from the *Tara* Oceans Interactome (Lima-Mendez et al., 2015) and the relative proportion of positive (green) and negative (red) interactions was computed for each major taxonomic group and plotted by decreasing negative edges. The absolute number of interactions between the taxonomic group and radiolaria is given in the legend. Diatoms are the 12th group from the left.

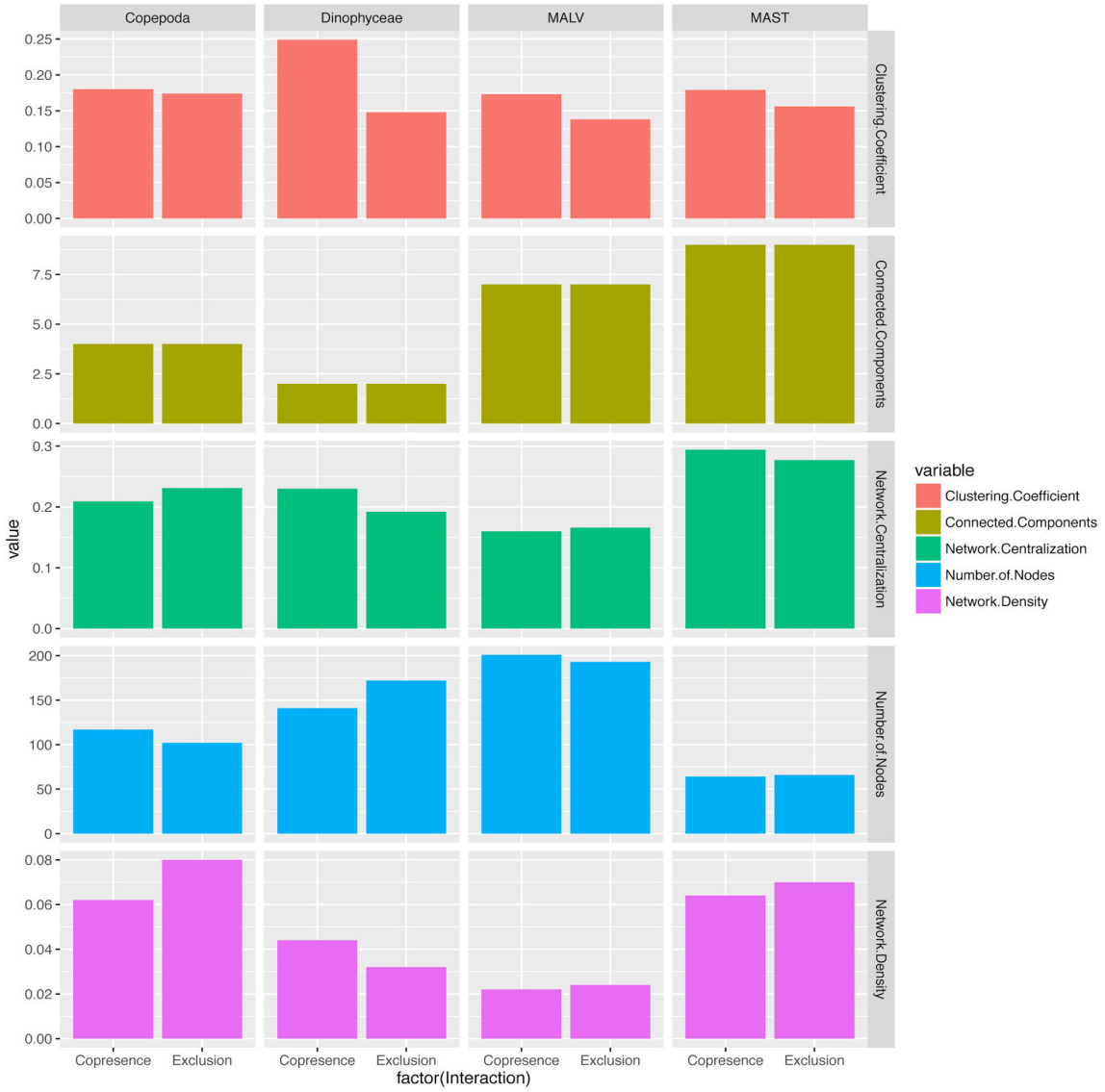


Figure S3.7. Subnetwork descriptors for major groups interacting with diatoms.
The diatom network was divided in sub networks based on the taxonomic affiliation of the nodes. For each subnetworks, relevant metrics of network topology were calculated under Cytoscape, for both copresence and exclusion subnetworks. The colours correspond to the measures, the taxonomic groups of the subnetwork is on the top part of the figure, and the type of interaction (positive or negative) in the bottom part of the figure.

Figure S3.8. Major diatom groups involved in the *Tara* Oceans interactome.
KRONA plot of the most important diatoms in the *Tara* Oceans Interactome based on the taxonomic affiliation of nodes. For example, 2% of the diatom interactions involved mutual exclusion between *Leptocylindrus* and another organism. A total of 81 unique diatom nodes (md5sum) were involved in the interactome.

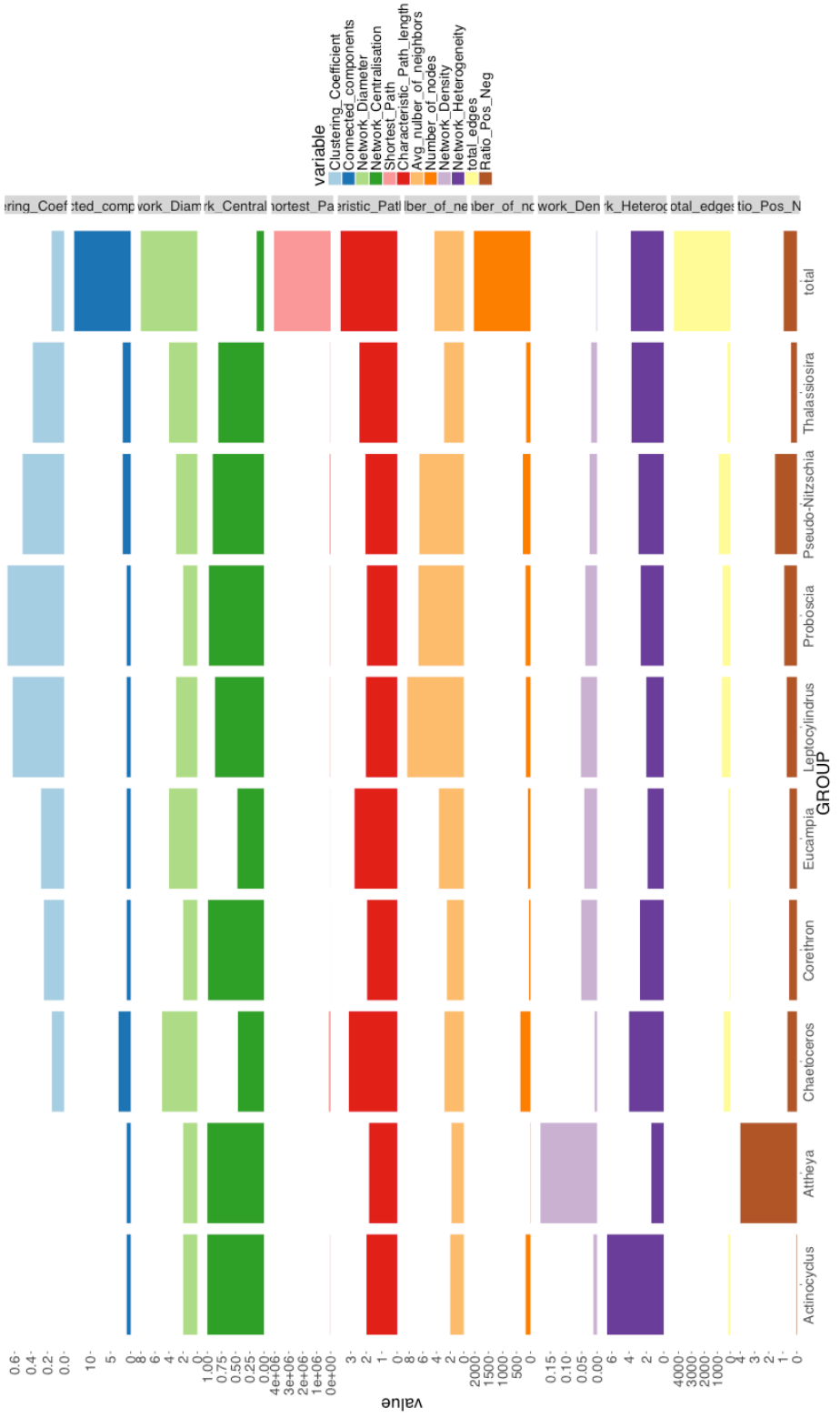


Figure S3.9. Subnetwork topologies of the top 10 most connected diatoms.

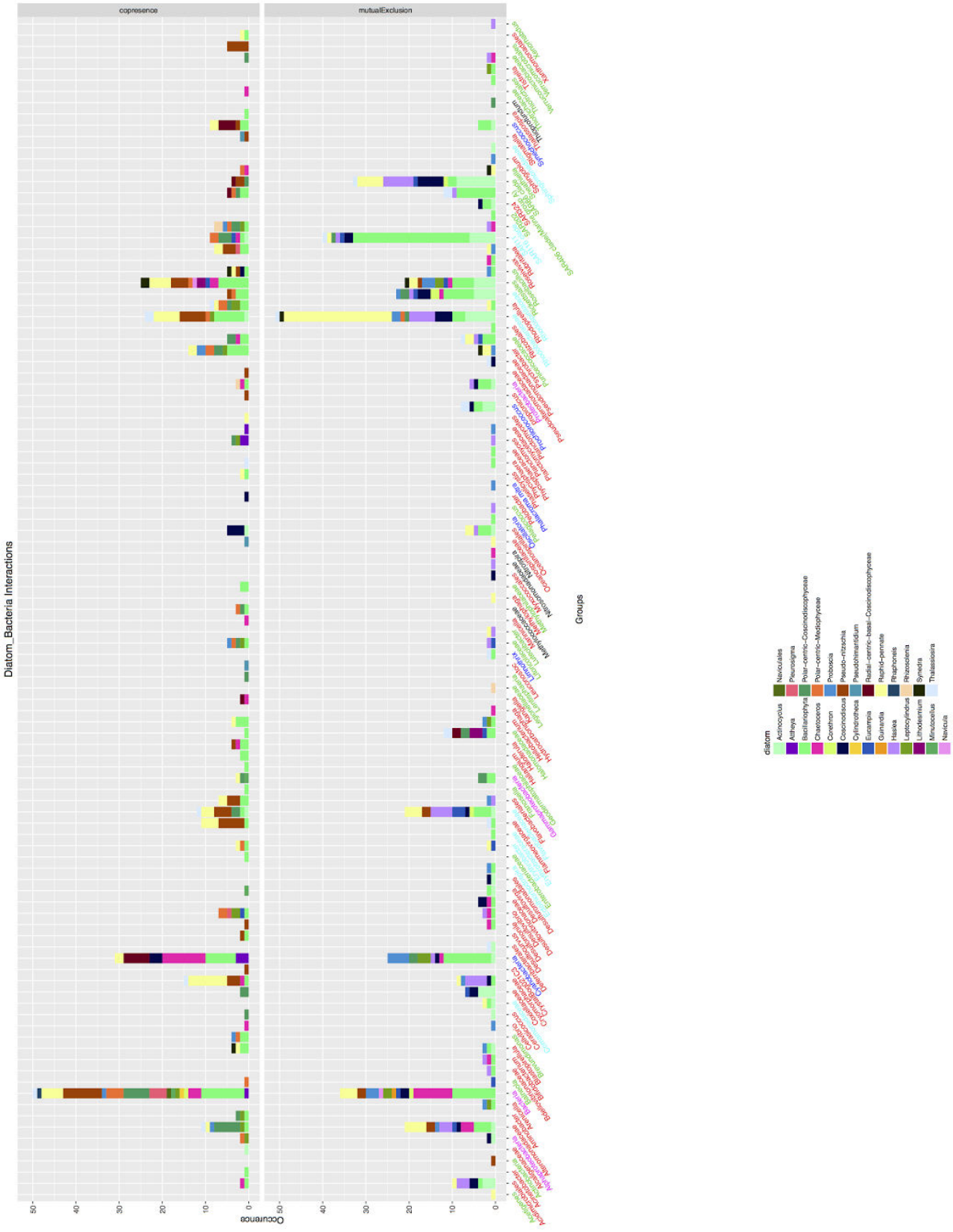


Figure S3.10. Distribution of diatom - bacteria interactions in the open ocean.
Diatom-bacteria interactions were derived from the global interactome. Bacteria involved are listed in abscissa and colored by trophic mode (Pink: unknown; Green:heterotroph; Red: chemoheterotroph; Blue: photoautotroph; Cyan: photoheterotroph; Black: chemoautotroph). The number of interactions in which they are involved in is represented by the coloured bars.

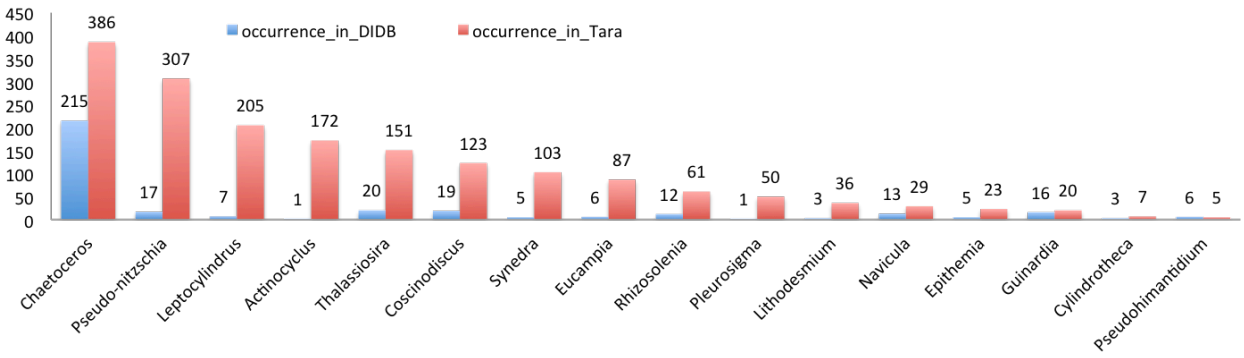


Figure S3.11. Comparison of diatom occurrence in the literature and in *Tara* Oceans interactome .
The interactome contains 32 unique genera of diatoms; 17 are completely absent from the literature (including poorly assigned ribotypes).

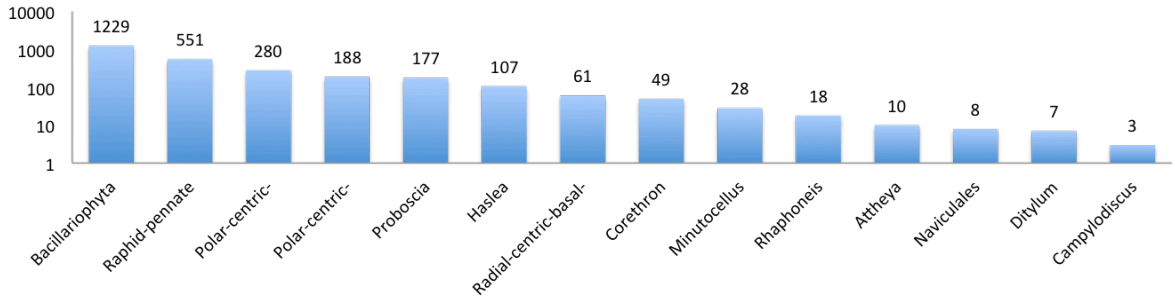


Figure S3.12. Understudied important interactors.
Many diatoms that play an important role in the network are absent from the literature, such as *Proboscia* (sub arctic diatom). The number indicates the number of edges in the Interactome.

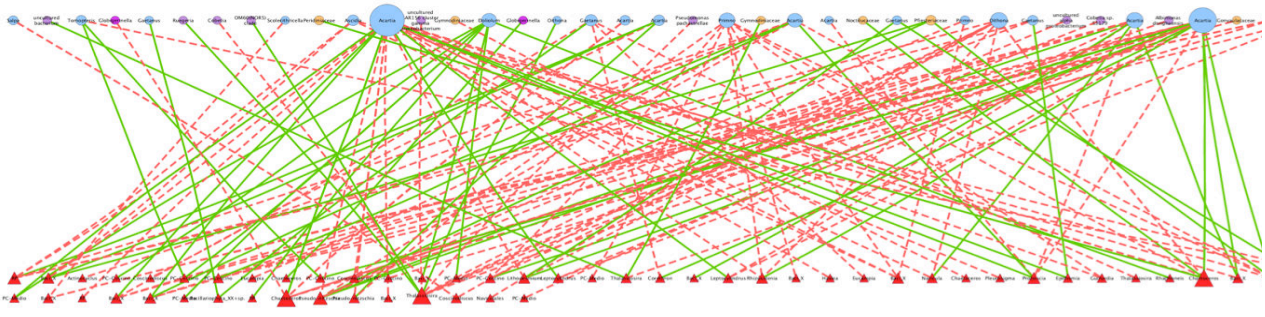


Figure S3.13. Predation pressure on diatoms in the open ocean.
Selection of *Tara* Oceans interactions confirmed by the literature, labeled as “eatenBy“. Diatoms are red triangles at the bottom part of the graph; potential predators are on the upper part of the graph, colored by their taxonomic group. Green filled lines correspond to positive edges, and red dotted lines correspond to negative edges.

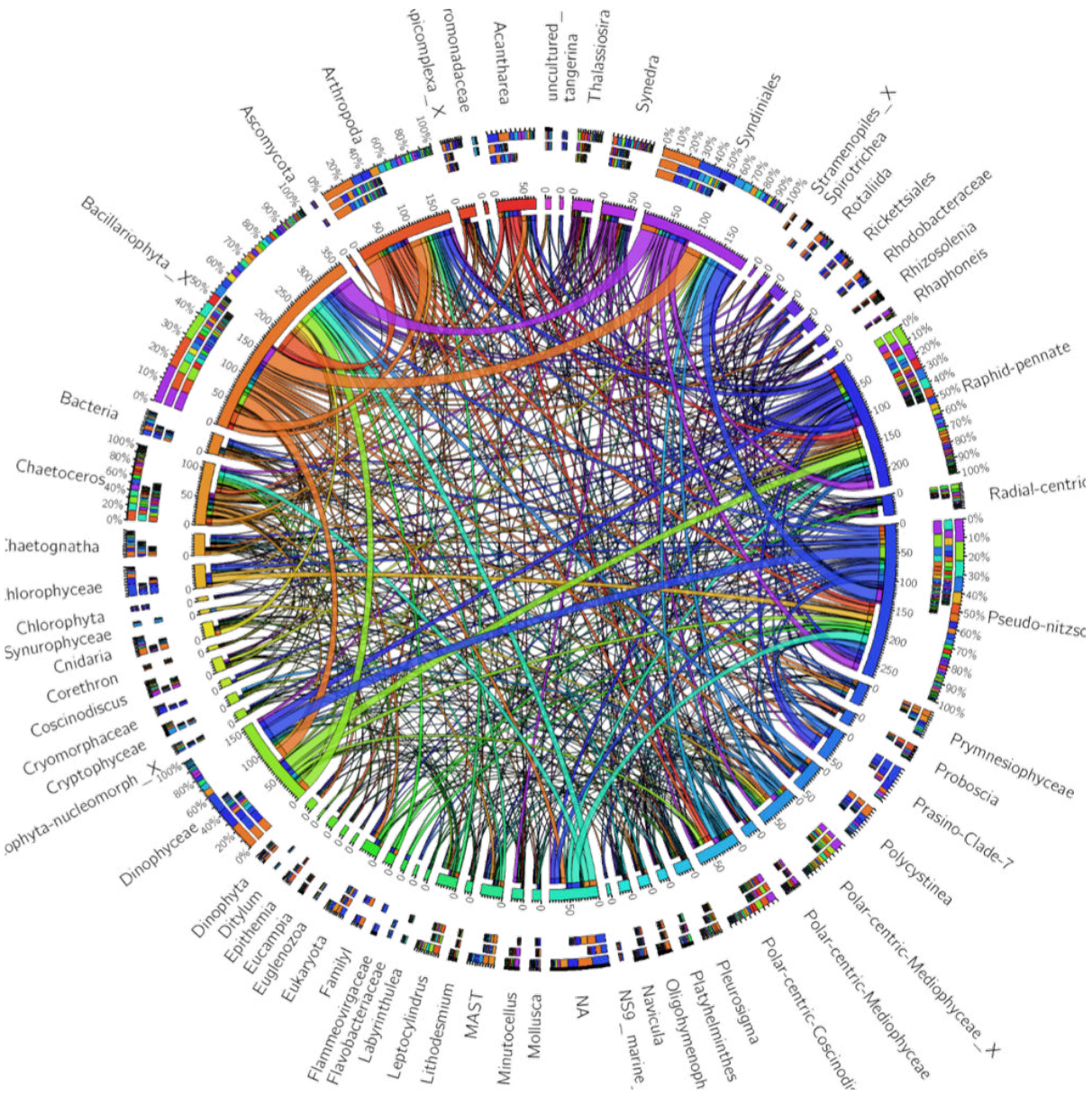


Figure S3.14. Diatom copresence (positive correlations) in the open ocean.
CIRCOS plot of diatom positive correlations in the *Tara* Oceans Interactome. Each node was grouped by its labeling (external ring with taxonomic affiliation), and is assigned to a color shown in the inner circle. The inner circle also provides the number of interactions for each taxonomy (Bacillariophyta, in orange, are involved in 350 interactions). The three external circles with stacked bars correspond to incoming (inner), outgoing (middle), and total (external) proportion of edges relative to partners of interaction (In total, Bacillariophyta have 17% of their total interactions with Syndiniales in purple).

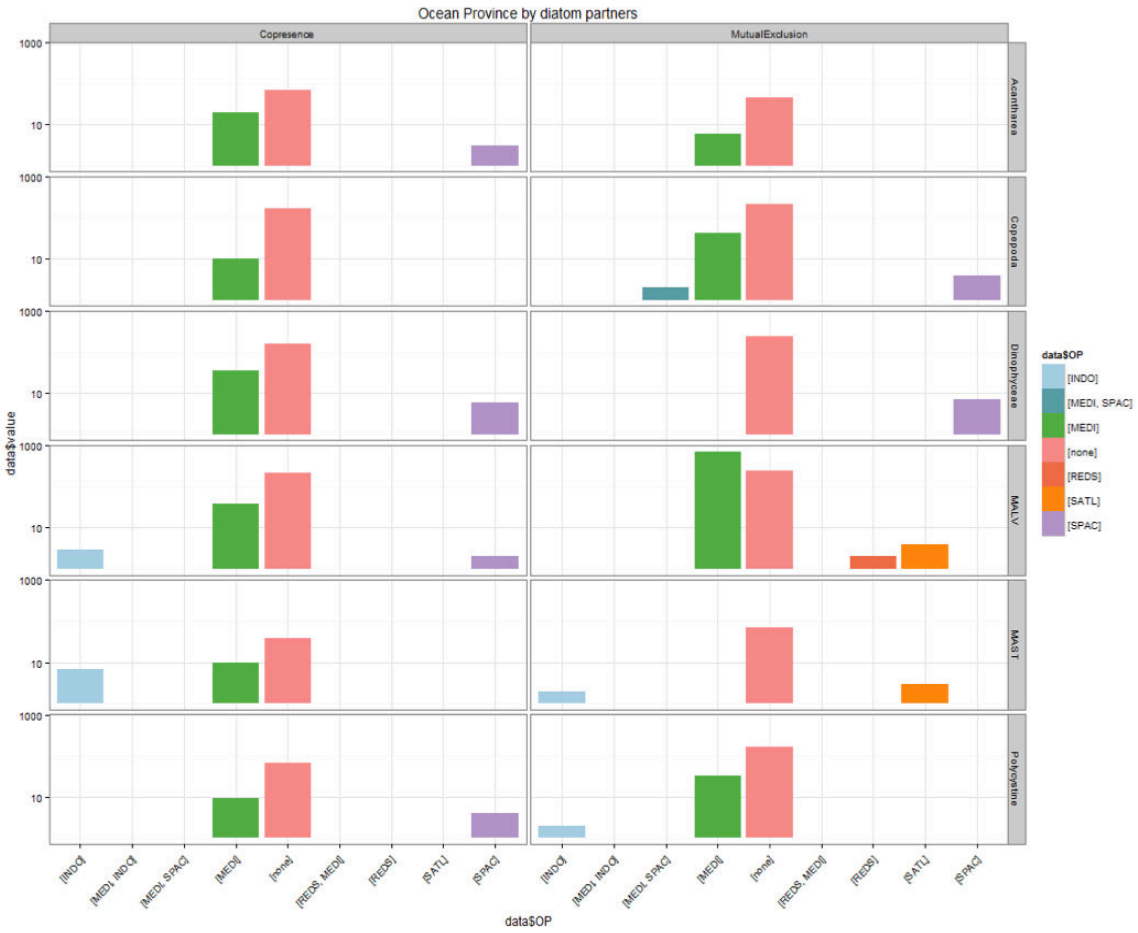


Figure S3.15. Ocean province significantly drives the observed interactions.
For each edge in the *Tara* Oceans Interactome, the importance of sample from a specific ocean province was assessed (Lima-Mendez et al., 2015). Here, we see which ocean province are significant for diatom interaction in sub networks of major interacting partners. Indian Ocean [INDO], Mediterranean Sea [MEDI], no driving ocean province [none], Red Sea [REDS], South Atlantic Ocean [SATL], South Pacific Ocean [SPAC]

Chapter 4: Characterisation of an abundant and widespread interaction

Summary

Characterisation of an abundant and widespread interaction between the pennate diatom *Fragilariopsis doliolus* and tintinnids

Flora VINCENT[1], Sébastien COLIN[2], Sarah ROMAC[2], Eleonora SCALCO[3], Lucie BITTNER[4], John R. DOLAN[5], Adriana ZINGONE[4], Colomban de VARGAS[2], Chris BOWLER[1]*

[1] Institut de Biologie de l'École Normale Supérieure, École Normale Supérieure, Paris Sciences et Lettres Research University, CNRS UMR 8197, INSERM U1024, F-75005 Paris, France

[2] Station Biologique de Roscoff, CNRS, UMR 7144, Place Georges Teissier, 29680 Roscoff, France

[3] Stazione Zoologica Anton Dohrn, Villa Comunale, Naples, 80121, Italy

[4] 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

[5] Sorbonne Universités, UPMC Univ. Paris 06, CNRS, Laboratoire d'Océanographie de Villefranche, CNRS UMR 7093, 06230 Villefranche-sur-Mer, France

* Author for correspondence; cbowler@biologie.ens.fr

Abstract

Diatoms are a diverse and ecologically important group of phytoplankton. Although most species are exclusively free-living many are found in biotic interactions with other organisms such as heterotrophic plankton. However, detailed molecular and morphological characterization of any such partnership is lacking so far, and an appraisal of the large-scale distribution and ecology of interactions has never been attempted. Through data-driven approaches, here we characterize, on a global scale, the epibiotic association between *Fragilariopsis doliolus*, the fourth most abundant Raphid-pennate diatom in the ocean for which we report the first molecular data, and tintinnid ciliates. Despite the diversity of ciliates involved, no mechanical adaptation to attachement was observed. Both partners displayed polymorphic sites in their V9 subregion with respect to available reference sequences, concomitant with ITS+5.8S+28S based haplotypes. Partial Least Square analysis using the *Tara* Oceans data shows that diatoms and tintinnids involved in these interactions differed in their principal abiotic explanatory variables and were associated in opposite relationships with temperature. The fact the organisms co-occur in nutrient rich samples, despite distinct relationships to other abiotic factors, illustrates the importance of studying biotic interactions to understand how they structure the community from the single cell to the global ocean.

4.1. Introduction

Phytoplankton are photosynthetic marine microbes that are responsible for half of Earth's net primary production (Field, 1998). Diatoms, a ubiquitous and predominant component of phytoplankton enveloped in a characteristic silica cell wall, are believed to be responsible for approximately 40% of marine net primary productivity (Nelson et al., 1995). They serve as the basis of the marine food web and are significant players in global biogeochemical cycles, representing a key unit of the marine ecosystem (Falkowski, 2002; Smetacek, 1998). Diatoms are frequently reported to dominate phytoplankton communities in well-mixed coastal as well as upwelling regions, where light, iron, silica, phosphorus, and nitrogen are available (Morel, 2003) however they can be constrained by nutrient availability in the open ocean (Cullen, 1991). They are also frequent and diverse in open ocean oligotrophic systems

(Malviya et al., 2016) where their survival in such low-nutrient regions is sometimes dependent on the presence of heterocystous N2-fixing cyanobacteria known as diazotrophs, which live in obligate or facultative symbioses with them (Foster et al., 2006). These associations are cases of mutualistic symbiosis, wherein diazotrophic bacteria like *Richelia intracellularis* and *Calothrix rhizololeniae* provide nitrogen in usable forms to diatoms including *Hemiaulus* spp and *Rhizosolenia* sp., which in return serve as protective hosts. This strategy to survive or adapt to a planktonic lifestyle is one example of the high diversity of biological interactions involved in the structure and functioning of the marine ecosystem, being likely the result of millions of years of co-existence and co-evolution of these planktonic organisms (Kiørboe, 2008).

Planktonic diatoms have been described in numerous other biological interactions, involving a range of organisms across all domains of life, as well as viruses. Beyond predation and competition, examples include host-parasite associations (e.g., between the chytrid parasites *Zygorhizidium planktonicum* and *Rhyzidium planktonicum* with the spring bloom diatom *Asterionella formosa* (Gsell et al., 2013)), various synergistic and parasitic interactions with bacteria (Amin et al., 2012; Sison-Mangus et al., 2014), endosymbiotic diatoms in nummulitid foraminifera (e.g., *Thalassionema* sp. related sequences in *Heterostegina depressa* foraminifera, (Holzmann et al., 2006)) and dinoflagellates (e.g., *Galeidinium rugatum* and *Durinskia baltica* (Schnepf & Elbrächter, 1999; Takano et al., 2008)) or even less understood forms of physical attachement with bacteria (Kaczmarska et al., 2005), copepods (Gárate-Lizárraga et al., 2009 ; Fernandes et al., 2012), diatoms (e.g., *Pseudo-nitzschia linea* and *Chaetoceros* sp. (Ruggiero et al., 2015)), *Phaeocystis* colonies (Sazhin et al., 2007), flagellated stramenopiles (Gómez et al., 2011), vorticellids (Nagasawa et al., 1996) and tintinnids (Gómez et al., 2007; Froneman et al., 1998). The latter examples, are representative of an interspecific association known as epibiosis (from the greek *epi* "on top" and *bios* "life"), which result in "spatially close associations between two or more living organisms belonging to the same or different species" (Harder, 2009).

Tintinnids (Choreotrichida) are heterotrophic planktonic ciliates enveloped in a species-specific test composed of organic material, the lorica (Agatha, 2013). They represent one of the most morphologically diverse groups of planktonic protists (Bachy, 2013). They are abundant and ubiquitous, commonly found in marine surface waters in concentrations ranging from $10^{\wedge 1}$ to $10^{\wedge 4}$ cells per liter (McManus & Santoferrara, 2013). Various associations between tintinnids and diatoms have been reported in the literature even though the true nature of the association remains unknown (Decelle et al., 2015) such as the one involving *Chaetoceros* sp. and *Eutintinnus* sp. known for over a century (Fol, 1883; Pavillard, 1913). With regards to *Fragilariopsis* sp. there have been occasional reports of occurrences with tintinnids. It was found with *Eutintinnus* sp. in material from near the Galapagos (Pavillard, 1935) and the Central South Pacific (Balech, 1962). More recently, *Fragilariopsis doliolus* was found associated with *Salpingella subconica* in material from the Southern Ocean near Prince Edward Islands with rates of occurrence close to 30% of all *Fragilariopsis* sp. encountered (Froneman et al., 1998). Of a different nature would be the association of *Laackmanniella* and other tintinnids with empty frustules of *Fragilariopsis* spp. and several other diatoms covering their lorica in the Antarctic, for which it was hypothesized that the ciliates retain diatom frustules following ingestion of the cellular contents (Gowing et al., 1992; Wasik et al., 1996;). However, a detailed molecular and morphological characterization of the partners is lacking so far, while an appraisal of the large-scale distribution and ecology of these consortia has never been attempted.

The recent *Tara* Oceans worldwide expedition has generated an immense amount of multidisciplinary information focusing on open ocean plankton communities in the surface ocean (Karsenti et al., 2011; Bork et al., 2015), enabling data-driven studies of biotic associations to emerge, using high-throughput genomics coupled with bioinformatics, as well as with high-content microscopic analysis of conserved fixed samples (Lima-Mendez et al., 2015; Mordret et al., 2015). One of the analytical platforms of the *Tara* Oceans project involves the high-throughput analysis by confocal laser scanning microscopy of formol-glutaraldehyde-fixed fractions enriched in nano- and micro-plankton (CLSM; Karsenti et al., 2011; see Methods). Analysis of samples from station 66 in the Benguela current (off South Africa) led to the initial observation of high abundances of the *Fragilariopsis doliolus* –

Salpingella sp. consortia (**Figure S4.1 & Figure S4.2**). We identified the V9 subregion of the 18S rDNA for both partners in the consortia and traced their worldwide abundance across the *Tara* Oceans metabarcoding data set, revealing eleven stations in which both organisms were highly abundant (see Methods). By examining those new samples through single cell microscopy, we investigated the morphological specificity of the consortia and whether mechanical adaptations permitted the physical attachement of the two partners. Through single cell sequencing, we addressed the phylogenetic diversity of the organisms including if the morphological features and geographic location of the partners involved had a signature at the genetic level. Finally, we examined in which environmental conditions the association occurred, hypothesing that the gain of motility for diatoms and protection for the ciliates was likely to happen in predator abundant nutrient rich samples.

4.1. Results

4.1.1. Morphological diversity of diatom-tintinnid consortia

Diatom-associated ciliates displayed diverse morphologies, of at least four different species of tintinnid ciliates. **Figure 4.1 (a-b)** shows the typical smooth lorica and trumpet shaped oral opening of *Salpingella acuminata* pictured in station 137, whereas **Figure 4.1 (e-f)** displays the less-differentiated oral end and ridged lorica joined in the aboral end, typical of *Salpingella faurei* extracted from station 66. In the consortia, the diatom was also identified along with *Salpingella curta* in Station 102 (not shown) and *Eutintinnus* sp. found in Station 124 on **Figure 4.1 (c-d)**. Morphological features allowed unambiguous and systematic identification of the diatom *Fragilariopsis doliolus*. In girdle view, rectangular shaped cells are united into curved ribbons by the valve surface (**Figure 4.1 (g)**). Valves are semi-lanceolate with bluntly rounded ends, with one straight margin and one broadly curved margin. (Wallich, 1860). Chloroplasts are present on either side of the median trans-apical plane and located close to the convex side of the valve (**Figure S4.1 (c)**). Scanning electron microscopy (SEM) of the consortia from Station 102 (**Figure 4.2 (a)**) in the Peruvian upwelling revealed a diatom displaying transverse striae with two alternating rows of poroids, confirming the species assignation to *Fragilariopsis doliolus* (Medlin and Sims, 1993).

Several features of the consortia are noteworthy. Contrary to other cases of diatom - tintinnid associations in which lorica are coated with diatom empty frustules, the diatoms had intact cell contents (**Figure S4.1 (b)**). The tintinnid also showed the ciliate cell within the test, with the exception of *Eutintinnus*, in which it was most likely the fixative, lugol, formol or formol-glutaraldehyde, that induced cell loss of the ciliate, as it was repeatdly observed containing its cell under light microscope on ethanol fixed samples. Presumptive colony formation of *Fragilariopsis doliolus* was found, as we observed apparently different and progressive stages, from less than half of the tintinnid being surrounded (**Figure 4.1 (f)**) to a nearly complete barrel (**Figure 4.1 (g)**) totally surrounding the ciliate. The diatom - tintinnid size ratio differs strongly depending on the ciliate cell involved, differentiating cases in which a complete ring would either fully hide the ciliate whose lorica opening is generally located just at the end of the diatom cell (**Figure 4.1 (e-f-g-h-k-j)**), or only partially in which case the diatom barrel is located closer to the oral end than aboral end (**Figure 4.1 (a-b-c-d-i**). SEM observations also revealed details of the contact between the two organisms, which apparently consisted of membranelles along the surface of the ciliate test adhering to the intact diatom frustule (**Figure 4.2 (a-b)**). Conversely, three dimensional scanning reconstructions showed no evidence of further physical attachment structures nor adaptation (**Figure 4.2 (c-d-e)**).

4.1.2. Phylogenetic identification of the interacting partners

A total of 25 *Fragilariopsis doliolus* – *Salpingella* spp. individual consortia were isolated by advanced micromanipulation from eleven stations identified through a data driven approach (see Methods). Following DNA extraction, group specific PCR amplification was performed to sequence the 18S, ITS, 5.8S, and partial 28S molecular markers for both the diatom and the ciliate. Sequences were obtained for at least one of the partners, yielding sequences for 21 diatoms and 17 tintinnids for which at least one of the markers was obtained (**Table 4.1**). We genetically identified both partners for 17 different consortia originating from Stations 66, 70, 102, 106, 122 and 124.

Phylogenetic analysis of the tintinnid 18S rDNA confirmed the presence of *Salpingella* specimens, as all sequences grouped in a monophyletic clade with strong support with all other sequences of the genus (BV 97,6% PP 1). The tintinnid sequences obtained here all grouped together (with low support (BV 61,5% PP 0,56), however the topology and the branch length suggest they might correspond to the same phylogenetic species. We were not able to identify *Salpingella faurei*, *Salpingella acuminata* and *Salpingella curta* at the genetic level in our sequences, despite their reported description in the consortia, or if the sequences correspond to a new species due to the limited resolution of the available sequence information (**Figure 4.3 (a)**). None of the *Fragilariopsis* – *Eutintinnus* isolated consortia were successfully sequenced. The ITS and 28S markers were chosen to perform the diatom phylogeny due to the fact that 18S rDNA is a relatively poor informative marker to distinguish *Fragilariopsis* species. The diatom sequences of the isolated specimens all branched with and within the *Fragilariopsis* genus (BV 91%) and formed a monophyletic group (BV 96%) (**Figure 4.3 (b)**). At this date (22/07/2016) these are the first publicly available and assigned sequences of the diatom *Fragilariopsis doliolus*.

To estimate population divergence among sequences for both taxonomic groups, a finer analysis to estimate divergence among sequences was performed using the ITS1&2, 5.8S and partial 28S sequences (D1 region). The tintinnid sequences reveal a total of eight different haplotypes based on 17 sequences of 721 base pairs length with two major haplotypes composed of sequences originating from stations belonging to distant provinces (**Figure 4.4 (a)**). No morphological differentiation corresponding to haplotypes was observed, although haplotype 1 generally corresponded to tintinnids with a short-looking lorica under light microscopy and haplotype 2 generally corresponded to a longer shaped morphotype. Overall, two tintinnid sequences (TI_823 and TI_884) displayed more than eight nucleotide differences with these two major haplotypes ; besides these the other tintinnid sequences diverged by less than five nucleotides over a 721 base pair alignment. The diatom ITS+5.8S+28S sequences showed 15 different haplotypes (**Figure 4.4 (b)**) from 20 sequences, which diverged up to nine nucleotides over a 727 base pair alignment. Diatoms originating from the South Atlantic region of Cape Agulhas (station 66) grouped all together compared to the rest of the sequences extracted from station 70 and from South

Pacific Ocean samples. The matching of the pairs revealed no specificity in combination of tintinnid haplotypes along with diatom haplotypes.

The sequence of the V9 18S rDNA sub-region (Amaral-Zettler et al., 2009) was trimmed from the amplified contigs for both partners of the consortia. It was recovered for 21 diatoms, although six of them were partial (**Table S1**). Focusing on the region without missing data, three variable positions out of 130 base pairs were identified with respect to other V9 sequences of *Fragilariopsis* species available on NCBI, in positions 58, 70 and 86. *Fragilariopsis* species therefore display a 98% similiarity for this molecular marker. All V9 sequences could be retrieved for the 17 sequenced tintinnids and were complete. Only one significant variable site was reported in position 55 and the sequences displayed a three base pair difference with the V9 sequence available on NCBI of a *Salpingella acuminata* isolate FG304. Interestingly, all tintinnids belonging to Haplotype 1 (except TI_887) displayed a T in position 55, and all tintinnids belonging to Haplotype 2 displayed a C in position 55. Initial cloning of 18S rDNA amplified from a single clonal diatom chain also revealed individual genetic micro diversity, with at least three different copies of 18S containing 28 nucleotide differences on a 1,246 bp alignment The three 18S copies also happened to be different at the level of the V9 subregion as shown in **Table S4**, with a maximum distance of four nucleotides difference between two V9 subregions. A single 18S rDNA tintinnid sequence showed at least four different copies of 18S displaying four polymorphic sites. However, three of those copies had an identical V9 sequence.

4.1.3. Geographic distribution and ecological context of the interaction

Amongst the 15 complete diatom V9 sequences obtained, eleven matched to a 100% identity with the *Tara* Oceans metabarcode f2f8b6bc0f4f3b6be690e0bd0f740a20 (f2f8) ; one diatom V9 matched 100% identity with the barcode 53cf4eb045f0a758fc188f13fcd54504 (53cf), both of which were assigned to Raphid_pennate_X+sp. (i.e. the environmental sequence JQ782062.1.1788_U); three complete V9 from isolated diatoms did not match any barcode in *Tara* Oceans (**Table S1**). F2f8 represented the fourth most abundant unassigned diatom barcode, with 301,093 reads over the 293 samples analysed in (Malviya et al., 2016), representing 24,5% of the

abundance of all the Raphid-pennates_X+sp in the top 100 unassigned barcodes. This single barcode, is nearly as abundant as all barcodes assigned to the *Pseudo-nitzschia* genera (down to 80% of sequence identity) grouped together that represents the 7th most abundant diatom genera with 305,115 reads, based on Malviya et al. In light of the extended *Tara* Oceans dataset (in prep), encompassing 150 stations, f2f8 is the second most prevalent *Fragilariopsis* sequence found in the open ocean with a total abundance of 414,113 reads. It is distributed across the whole *Tara* Oceans sampling expedition, with higher abundances reported below the Equator and occurrences both in coastal and open ocean stations (**Figure 4.5 (a)**). Comparatively, 53cf had a total abundance of 853 reads across the whole *Tara* Oceans data (**Figure S4.3 (c)**).

In total, the tintinnid V9 sequences matched at 100% identity to three unique barcodes in the *Tara* Oceans dataset: b61a7b36be9517d26828b362d8e147e1 (b61a), a7cbcf338bce632eeb1d94a010707449 (a7cbc) and deb2a198341ff3dcaf8e594767503905 (deb2a), all three of which were assigned to *Choreotrichia_XX+*sp (i.e. an environmental lineage GB GU819299.1.1181_U). A7cbc was particularily abundant and widely distributed, compared to b61a and deb2a, with a total abundance of 106,292 reads over the complete data set (**Figure 4.5 (b)**). B61a had a total abundance of 24,839 reads and deb2a of 3,660 reads, the second being restricted to the Benguela's upwelling current (**Figure S4.3 (a-b)**). All the tintinnids belonging to haplotype 1 (except TI_887) displayed a V9 corresponding to b61a, and all the tintinnids belonging to haplotype 2 displayed a V9 corresponding to a7cbc ; no particular pairing the the V9 sequences of the two partners was observed (**Table S2**).

Individual barcodes extracted from fractions 20-180 and 180-2000 micron of the surface layer were analysed with respect to the most ecologically relevant oceanographic variables, such as temperature, salinity, phosphate and silica through partial least square analysis (**Figure 4.6**). F2f8 and 53cf barcode abundances in surface samples for fractions 20-180 microns and 180-2000 microns were best explained by temperature (regression coefficient 0.38 and 0.019 respectively), nitrate (0.35 and 0.014) and density (-0.28 and -0.014). Tintinnid barcodes displayed different explanatory variables, but abundances of a7cbc and b61a were best explained by temperature with a negative (-0,4) and positive (0,22) regression coefficient respectively, followed by density (0,27) and chlorophyll (0,25) for

a7cbc, chlorophyll (-0,18) and silica (-0,15) for b61a (**Table S7c**). F2f8 (diatom) and a7cbc (tintinnid) display an antagonitic response to their major abiotic explanatory variable (temperature), yet the organisms are found paired together in similar samples.

A common feature amongst the majority of samples in which the consortia was observed is their proximity to nutrient rich regions illustrated by, in general, higher nitrite concentrations as shown in **Figure S4.4** but also oceanic province. For example, station 66 is located in the Benguela upwelling current, station 70 is situated at the limit of the nitrate plume originating from the Benguala current flowing northwest, stations 102, 106 and 109 are in the Peru current, stations 122 and 124 benefit from the island mass effect of the Marquesas and stations 137 and 138 were sampled in the Californian current.

Beyond abiotic variables, co-occurrence of the two partners and other organisms of the plankton was explored, as tintinnid competitors such as oligotrichs and potential predators such as copepods are known to impact tintinnid distribution (Dolan et al., 2002). Their total barcode abundance was assessed in the 150 stations sampled during the Tara Oceans expedition and compared with the occurrence of the interaction (**Figure S4.5**). Barcode abundance of potential competitors did not appear to be particularly relevant as it did not display an overlapping pattern with the diatom-tintinnid association; however, copepod abundance in fractions 20-180 and 180-2000 microns was found to be high in samples in which the association was found. We investigated the occurrence of the molecular sequences from our isolated interaction within the global ocean interactome (Lima-Mendez et al., 2015). One tintinnid barcode appeared and was a7cbc (**Figure S4.6**), which co-occurred significantly with *Eutintinnus* (found in association with *Fragilariopsis* in our samples) and *Amphorides* ciliates. Both diatom barcodes positively correlated with *Salpingella* were assigned to *Bacillariophyta_X* and *Ditylum*, and were highly divergent from *Fragilariopsis doliolus* based on the V9 sequence. Our conclusion is that a co-occurrence network at a large scale was inefficient in recovering this biotic interaction.

Finally, the ecological importance of the diatom-tintinnid association was quantified through actual cell counts in four different samples based on the same data driven process (**Figure**

4.7). The results reveal that diatom-associated *Salpingella* can represent up to 93.5% of all *Salpingella* present in a sample and constitute over 50% of the total tintinnid community (**Figure S4.7**) with 600 diatom-tintinnid associations interactions per ml of net sample from fraction 20-180 µm equivalent, to ~60 interactions/L, such as in station 102 located in the upwelling Peruvian west coastal current. The consortia was also observed off in the North Pacific Equatorial Countercurrent (Station 139), with approximately 200 interactions per ml of net sample from fraction 20-180 µm equivalent to ~20 interactions/L. Interaction counts between lugol- and formol-fixed samples showed high discrepancy, the association typically being much represented in acidic lugol, which is not effective in preserving colonies either. Fixative bias is well documented for tintinnid cells (Modigh, 2005), but also appears to affect the detection of biotic interactions.

4.1. Discussion

Using a data-driven approach enabled us to efficiently target individual samples for further study and characterize at a global scale the epibiotic association between *Fragilariopsis doliolus* and tintinnid ciliates. Microscopic analysis revealed high morphological diversity of tintinnids involved in the consortia, enveloped in a barrel shaped diatom chain. Different stages of the association were found, showing a tintinnid fully or partially encapsulated by *Fragilariopsis doliolus,* either due to an early stage of the association, or an unbalanced diatom/tintinnid length ratio. At each stage, the diatom and tintinnid displayed intact cell content that could signal a potential benefit of the association for both partners. No mechanical adaptation to attachment was observed, suggesting that the association is maintained by the secretion of adhesive extracellular polymeric substances (EPS) by the diatom, as already described in other raphid pennate diatoms and reviewed in (Wetherbee, 1998), an idea reinforced by the quantitative discrepancy between formol and lugol fixed samples.

Single pair sequencing and 18S rDNA phylogeny obtained for 17 specimens confirmed the genus level assignation of the tintinnids, that grouped with and within *Salpingella* spp. reference sequences. The ITS+5.8S+28S tintinnid revealed two major haplotypes were concomitant with a unique polymorphic site in the V9 subregion of the 18S rDNA marker,

showing Haplotype 2 as more abundant than Haplotype 1. However, these markers were not sufficient to identify morphologically characterized tintinnids. Diatom sequences formed a monophyletic clade strongly supported based on ITS+5.8S+28S alignement, confirming the good potential of those markers to study *Fragilariopsis* species (notably in comparison to the 18S rDNA sequences). However, the V9 subregion showed three polymorphic sites when compared to other *Fragilariopsis* species available on NCBI. A classic 97% identity clustering of OTUs from a metabarcoding survey would have certainly grouped together V9 sequences from *Fragilariopsis doliolus* and *Fragilariopsis curta* or *Fragilariopsis cylindrus*, yet none of the two last species could possibly interact the way *Fragilariopsis doliolus* does.

The biogeography of both partners was recovered by identifying the V9 in the *Tara* Oceans global metabarcoding data set. *Fragilariopsis doliolus* is the fourth most abundant Raphid-Pennate in *Tara* Oceans, occurring in both North and South hemispheres, coastal and open ocean regions as well as eutrophic, and oligotrophic waters. A similar distribution was observed for barcodes assigned to *Salpingella* with an additional occurrence in the Mediterranean Sea, suggesting that when both partners co-occur, they associate to form the described consortia. The most abundant barcodes of diatoms and tintinnids differed in their principal abiotic explanatory variables and opposite preferences to temperature. The fact the organisms co-occur, despite antagonistic responses to abiotic drivers, illustrates the importance of studying biotic interactions to understand how they structure the community, as was already highlighted in (Lima-Mendez et al., 2015). Analysing the presence of other members of the community such as competitors and predators can provide insight, in our case showing the association was concomitant with the abundance of copepods that feed both on diatoms and tintinnids. This leads to the remaining missing link with respect to the study of the diatom epibiotic assemblage involving the ecological and evolutionary implications of this association.

Pennate diatoms and tintinnid assemblages have been reported periodically for decades, in a wide range of ecological contexts from sub-antarctic regions to equatorial waters, involving different species of tintinnids and diatoms. The respective benefits for each partner to undergo such an association are difficult to assess but a few advantages both for

the diatom and the tintinnid can be envisaged. Diatoms use nutrients from the surface layer and assimilate them through photosynthesis, which can lead to the production of exudates of particulate organic matter that tintinnids can potentially benefit from (Gügi et al., 2015). It should also be noted that tintinnids can feed on diatoms (Gowing et al., 1992; Verity et al., 1986). In parallel, the association with diatoms could lower the tintinnid's susceptibility to predation by enveloping the ciliate in an "armour", and increasing the size, as the association occurs in regions where the pressure of potential predation is high. A further hypothesis is that attached ciliates see changes in fluid dynamics of the feeding current, that leads to steeper velocity gradients and higher flow rates close to the lorica (Jonsson et al., 2004).

In return, diatoms may benefit from increased motility by attaching to tintinnids, with potential higher access to nutrients or maintenance in the euphotic layer. However, this last hypothesis is weakened by the occurrence of diatom-tintinnid associations in resource plenty regions. Finally, it is interesting to note that few truly planktonic raphid pennates exist, *Pseudo-nitzschia* and *Fragilariopsis* amongst the most cited ones, and that the epiphytic lifestyle of some araphid pennate diatoms is considered as an ancestral stage . Most of them form chains and colonies in the plankton, a trait that has emerged over and over again in the life history of pennate diatoms (Kooistra et al., 2009). Reasons that have been advanced to underlie the low incidence of planktonic pennate diatoms is their sexual reproduction mode involving amoeboid gametes, for which encounter rates would be expected to be decreased in planktonic environments (Falkowski et al., 2007). Planktonic lineages have overcome this constraint by sinking to the bottom, forming blooms, or aggregating at the surface. Here, *Fragilariopsis doliolus* demonstrates its capacity to adopt an ancestral lifestyle tailored to the organisms co-occurring with it, potentially increasing their encounter rate. Such a hypothesis could implicate epibiosis as a supplementary strategy to warrant their ecological success.

Further studies are needed to elucidate the nature of the diatom-tintinnid association, although experimentation will be hampered by the absence of cultures of either partner. Bacterial communities from single and associated partners could be analysed, and modeling

of fluxes may enable investigation of biophysical aspects of the association. Transcriptome analysis may allow exploration of bioadhesion-related genes involved in cell adhesion or EPS synthesis. Our study not only characterizes a new diatom species of global ecological importance, but also highlights an abundant and ubiquitous marine microbial interaction by integrating data from the single cell to the global ocean.

4.1. Materials and methods

4.1.1. Morphological investigation of the consortium

The sampling strategy used in the Tara Oceans expedition is described in (Pesant et al., 2015), and samples used in the present study are listed in **Table S3**. The Tara Oceans nucleotide sequences are available at the European Nucleotide Archive (ENA) under the project PRJEB402 and PRJEB6610.

CLSM. Several consortia were analysed from formaldehyde-preserved samples and imaged with confocal laser scanning microscope (Leica TCS SP8), equipped with an HC PL APO 40 × /1.10 W motCORR CS2 objective. Multiple fluorescent dyes were used sequentially to observe the cellular components of the ciliate (host) and the microalgae, such as the nuclei (blue) and the cellular membranes (green) with Hoechst (Ex405/Em420-470) and DiOC6 (Ex488/Em500-520), respectively. Red autofluorescence of the chlorophyll (Ex638/Em680-700) was also visualized. Image processing and three-dimensional reconstructions were conducted with Fiji (Schindelin et al., 2012) and IMARIS (Bitplane) software.

4.1.2. Data-driven station selection for diatom-tintinnid consortium identification

PCR amplification of the V9 18S rDNA subregion. Single consortia composed of the association between the diatom and the tintinnid were initially isolated from formol-glutaraldehyde-fixed surface samples collected by a plankton net (20 -180 µm mesh-size) in Station 66 located in the Agulhas current, in which the association was initially observed as highly prevalent (**Figure S4.1**). Using a glass micropipette, interactions were rinsed two to three times in a minimum volume of sterile artificial sea water, before being immersed in

300 µL of Tissue and Cell lysis buffer from MasterPure DNA and RNA purification kit (Epicenter), and stored at -20°C. DNA extraction was done following the protocol of the MasterPure DNA and RNA purification kit (Epicenter). Polymerase chain reactions (PCR) using universal-eukaryotic primers V4F 5'- CCAGCASCTGCGGTAATTC – 3' and 15101R 5'- CCTTCYGCAGGTTCACCTAC- 3' were performed on total DNA extracts. PCR amplifications were conducted with the Phusion High-Fidelity DNA Polymerase (ThermoFisher Scientific). The PCR mixture (25 µL final volume) contained 2 µL of DNA, 0.5 µM (final concentration) of each primer, 3% of dimethyl sulfoxide, 200 µM dNTPs and 5X Phusion HF Buffer. Amplifications were conducted in the PCR with the following PCR program: initial denaturation step at 98°C for 30s, followed by 28 cycles of 10s at 98°C, 45s at 55°C, 15s at 72°C and final elongation step at 72°C. Amplicons were purified using the Nucleospin Gel and PCR Clean-up (Macherey-Nagel) with two NT3 rinsing due to the small quantities of DNA. In order to recover the partial 18S rDNA of both partners in the interaction, a cloning approach was adopted: purification was followed by a poly-A tailing reaction using GoTaq DNA Polymerase (Promega) and ligated in pGEM-T Easy Vector Systems (Promega). The product was cloned using chemically competent Escherichia coli cells, colonies were selected with blue white IPTG/Xgal and length of the insert was checked by PCR before picking. Plasmids were purified with the Plasmid DNA purification kit (Macherey-Nagel) and double-end sequenced by GATC Sanger sequencing.

V9 identification in the Tara Oceans data. Sequencing results were assembled and trimmed to retain the V4 to V9 18S fragment. The V9 fragments obtained from Station 66 were blasted against the Tara Oceans ribotype database and matched as the best result two different diatom barcodes (version of the V9), and three different tintinnid barcodes (**Table S4**). For each partner, one of the barcodes was particularly abundant and found present in the Tara Oceans metabarcoding global data set. Sequence f2f8b6bc0f4f3b6be690e0bd0f740a20 (f2f8) is assigned to Raphid-pennate_X+sp. and a7cbcf338bce632eeb1d94a010707449 (a7cbc) is assigned to Choreotrichia_X+sp. The search for these sequences in the global metabarcoding dataset revealed that both were widely distributed and displayed broadly overlapping patterns (**Figure 4.5**). Ten stations were chosen for additional analysis, based on the presence of at least 100 copies of each

barcode (f2f8 and a7cbc) in the sample : Stations 70, 102, 106, 109, 111, 122, 124, 128, 137, and 139, in fraction 20-180 micron in surface samples. We obtained morphological confirmation for the presence of the diatom Fragilariopsis doliolus associated with three different species of the tintinnid genus Salpingella and one species of Eutintinnus in all those stations. Abundance of the five different barcodes are available in **Table S5**.

4.1.3. Molecular and phylogenetic analysis of the diatom-tintinnid association

Advanced micromanipulation for cell isolation. Diatom-tintinnid consortia were isolated from ethanol-preserved samples collected by a plankton net (20 µm mesh-size) in the eleven Tara Oceans stations of the Indian, Atlantic and Pacific Oceans. By using an OLYMPUS IX51 inverted microscope equipped with an Eppendorf's manual microinjection CellTram® Air, single consortia were carefully rinsed exactly three times in 100% ethanol Labtech wells before being taken in picture, and then treated with a protocol for DNA extraction according to steps in the MasterPure DNA and RNA purification kit (Epicenter). Samples used in this study for molecular identification are available in **Table S1**.

PCR amplification of small subunit rRNA genes, internal transcribed spacers ITS1 and ITS2, and 5.8S rRNA genes. To obtain different phylogenetic ribosomal markers for both partners, initial group specific amplifications were conducted with the Phusion High-Fidelity DNA Polymerase (Finnzymes). Group specific primers were designed and inspired from (Bachy et al., 2013) and (McDonald et al., 2007) and are available in **Table S6**. The PCR mixture (25 µl final volume) contained 2 µl of DNA, 0.35 µM (final concentration) of each primer, 3% of dimethyl sulfoxide and 2× of GC buffer Phusion MasterMix (Finnzymes). Amplifications were conducted in a PCR thermocycler (Applied Biosystems) with the following PCR program: initial denaturation step at 98 °C for 30 s, followed by 37 cycles of 10 s at 98 °C, 30 s at the annealing temperature of 50°C 30 s at 72 °C and final elongation step at 72 °C for 10 min. Amplicons were purified using Nucleospin Gel and PCR Clean-up (Macherey-Nagel), reamplified using internal primers to cover the 18S, ITS1&2, 5.8S and partial 28S then Sanger sequenced using the ABI-PRISM Big Dye Terminator Sequencing kit (Applied

Biosystems). Amplicon sequences of both the diatom and the tintinnid were cleaned, trimmed, and assembled using Sequencher (version 5.4).

Phylogenetic analysis. For the tintinnid, contigs of the amplicons were obtained, and two matrixes of 18S rDNA and ITS+5.8S+28S rDNA were built, including reference sequences from (Bachy et al., 2012). For diatoms, similar matrixes were built including reference sequences from blast top hits in GenBank and reference sequences from (Theriot, 2010). The four matrices were completed with appropriate outgroup sequences, and aligned using MAFFT version 7 (Katoh et al., 2013). Sequences were trimmed using Gblocks (Talavera et al., 2007) as implemented in SeaView (Gouy et al., 2010) allowing for smaller final blocks, gap positions within the final block, and less strict flanking positions.with various degrees of stringent set selection. JmodelTest (Darriba et al., 2012) was used to determine the best model of nucleotide substitution for each matrix. The general time-reversible model with gamma distribution of rate variation (GTR+G) was selected for the diatom ITS+5.8S+28S, and the tintinnid 18S and ITS+28S tree. Phylogenetic inference was done using PhyML 3.0 (Guindon et al., 2010) and robustness of topologies assessed by performing 1000 bootstraps. Bayesian inference analyses were carried out on the tintinnid 18S rDNA tree, with the program MrBayes (Huelsenbeck, 2001), with two independent runs and 1,000,000 generations per run. After checking convergence (maximum difference between all bipartitions <0.01) and eliminating the first 3,500 trees (burn-in), a consensus tree was constructed sampling every 100 trees. Trees were visualized and edited using FigTree v.1.4.2. For ITS+5.8S+28S rDNA, a statistical parsimony networks was constructed with TCS software (95% connection limit, 10 or less connection steps, gaps considered as 5th state) (Clement et al., 2000), and visualised using TCS.

4.1.4. Environmental and community contextualization of the interaction

Partial Least Square Analysis. We investigated the environmental parameters explaining the abundance of the two partners. As variables were shown to be multicollinear, Partial Least Square Regression analysis was conducted on range transformed oceanographic data as predictors and Hellinger transformed abundance data of corresponding barcodes in the Tara Oceans data as responses using the « plsdepot » package in R version 3.3.0. The resulting

Circle of Correlation is displayed in **Figure 4.6** and regression coefficients are available in **Table S7c**.

Co-occurrence network. The five different V9 barcodes corresponding to isolated organisms were looked up for in the Tara Oceans interactome (Lima-Mendez et al., 2016), in order to characterize the organisms with which the barcodes were detected to be significantly correlated to, both positively and negatively. The network of co-occurrence is shown in **Figure S4.6** and correlation values are given in **Table S8**.

Cell quantification. To assess the ecological importance of the association in terms of abundance, cell/mL counts were performed in four Formol fixed samples as shown in **Figure 4.7**. Aliquots of formol fixed plankton net material (0.5 - 2.0 ml), equivalent to approximately 10 L whole water filtered, were examined in plankton settling chambers using an Inverted Olympus microscope (model IX71) equipped with DIC optics at 200x total magnification. Tintinnid identifications were made based on lorica characteristics and dimensions following standard taxonomic monographs (Kofoid, 1939; Hada, 1938). They were compared to cell/mL counts of total tintinnids encountered in other Tara Oceans samples as shown in **Figure S4.7** and raw counts are available in **Table S9**.

Acknowledgments

This work was supported by the project OCEANOMICS that has received funding from the French government, managed by the Agence Nationale de la Recherche, under the grant agreement 'Investissements d'Avenir' ANR-11- BTBR-0008. We thank the European Molecular Biology Laboratory (EMBL, Heidelberg) and more specifically Rainer Pepperkok for providing access to the advanced light microscopy facility (ALMF). We also thank the coordinators and members of the *Tara* Oceans expedition. We thank Charles Bachy for his precious help on tintinnid phylogeny, Thibaut Pollina for technical development of advanced micromanipulation and Yann Thomas for initial on help on molecular identification of the partners.

4.1. Figures and Supplementary Material

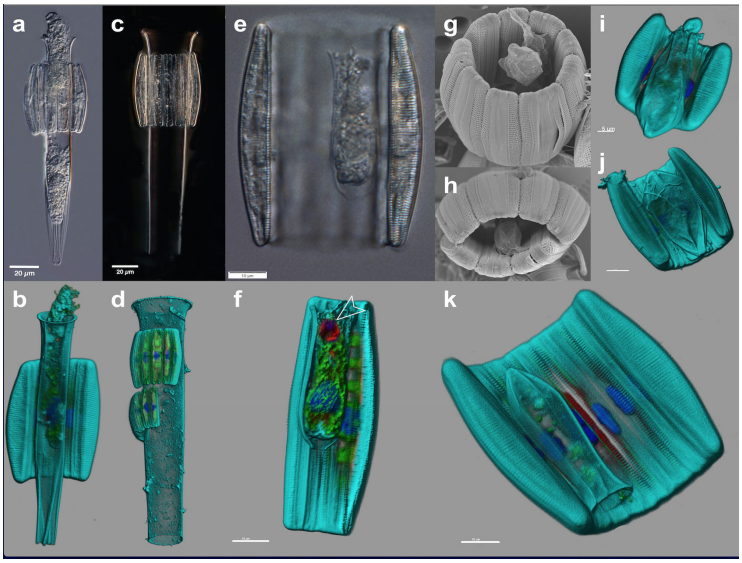


Figure 4.1. Diatom-tintinnid couples display high morphological diversity.
(a-b) *Salpingella acuminata* in Station 137. (c-d) *Eutintinnus fraknoi* with one or two diatom chains in Station 124. (e-f) *Salpingella faurei* in which CSLM reveals the presence of potential prey chloroplast (arrow) in the tintinnid lorica in Station 66. (g-h) Advanced status of diatom chain formation taken in Station 102. (i-j-k) Relative position of the lorica opening with respect to the diatom barrel.

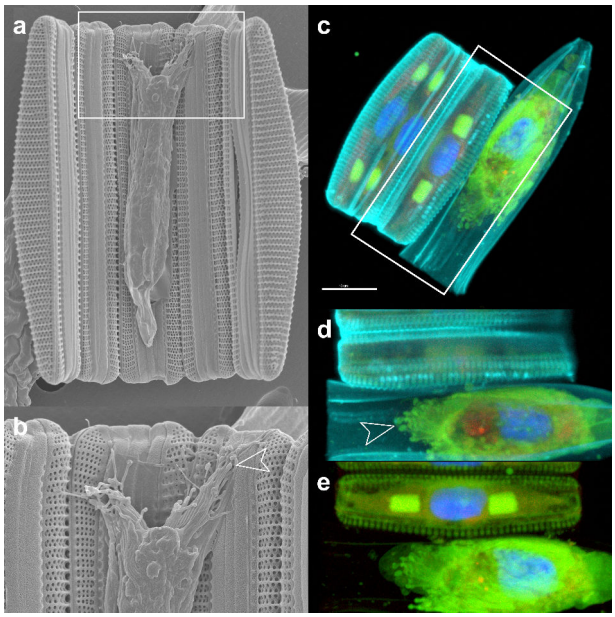


Figure 4.2. Close up view of the interface between the two organisms.
(a-b) Scanning electron microscopy in Station 102 and (c-d-e) CSLM.

Station	Ocean Province	Picture	N Single cell barcoding	Identifier "TI_"
66	Eastern Africa Coastal Prov.	Yes	6	882, 884, 886, 887, 888, 890
70	South Atlantic Gyral Prov.	Yes	1	881
102	Pacific Equatorial Divergence Prov.	Yes	5	820,821, 823, 825
106	Pacific Equatorial Divergence Prov.	Yes	4	851, 852, 853, 854
109	Pacific Equatorial Divergence Prov.	Yes	0	
111	South Pacific Subtropical Gyre	Yes	0	
122	South Pacific Subtropical Gyre	Yes	2	863, 864
124	South Pacific Subtropical Gyre	Yes	7	805, 807, 808, 811, 813, 814, 815
128	South Pacific Subtropical Gyre	Yes	0	
137	North Pacific Equatorial Countercurrent	Yes	0	
139	North Pacific Equatorial Countercurrent	Yes	0	

Table 4.1. Summary table of available information for each investigated station.
Each isolated interaction has a unique identifier starting with "TI_". Visual evidence was obtained either under light microscopy, confocal microscopy or scanning electron microscopy. For each single cell barcoding, at least one of the molecular markers (18S, ITS or partial 28S) was obtained.

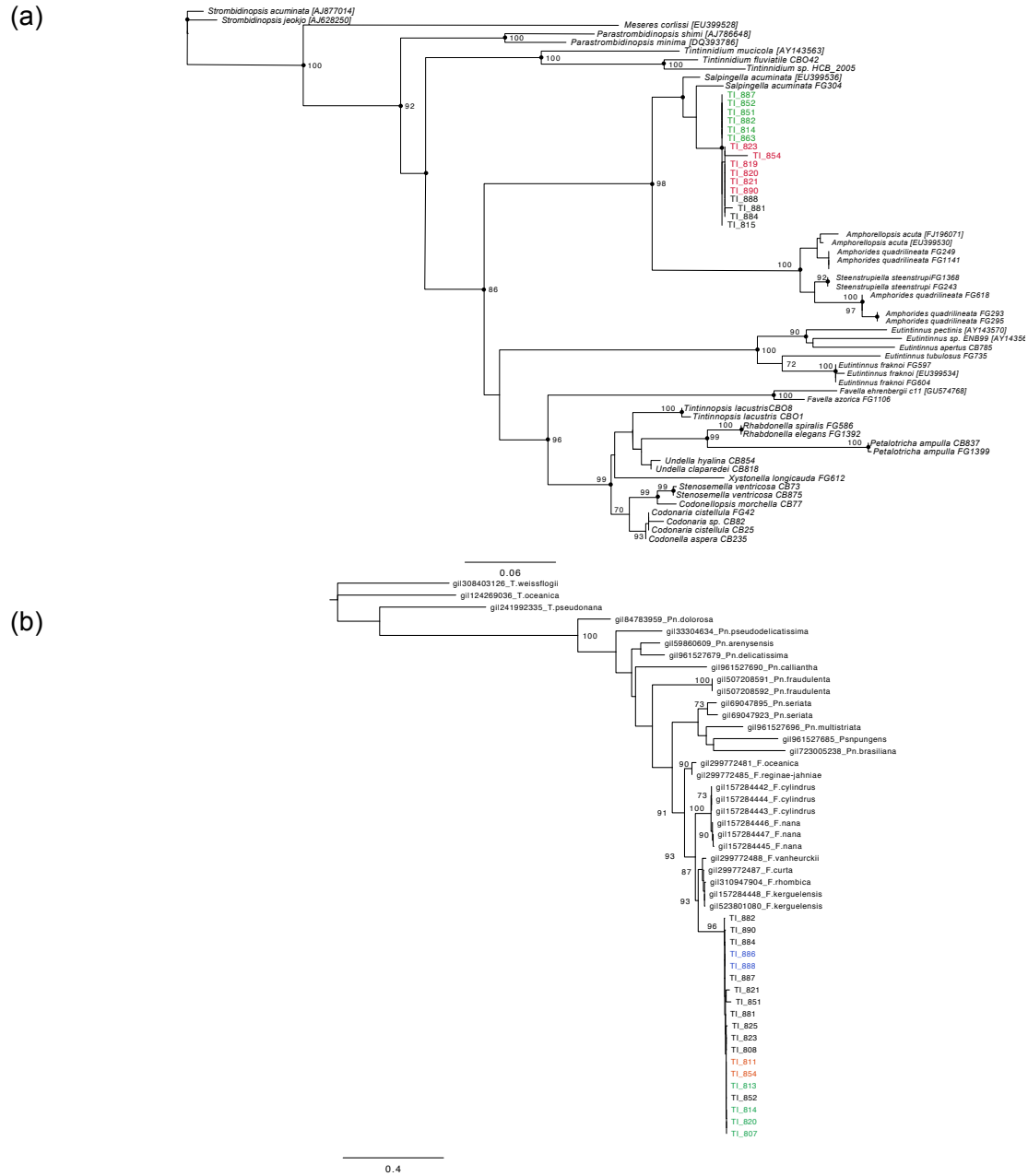


Figure 4.3. Phylogeny of the two partners.

(a) Maximum likelihood rooted phylogenetic tree of choreotrich SSU-rDNA, based on 1,455 aligned positions. Reference sequences were extracted from (Bachy et al., 2012). Numbers at nodes are bootstrap values in percentage rounded up for 1000 bootstraps (value below 70% are omitted). Bayesian posterior probabilities higher than 0.90 are indicated by filled circles. Accession numbers are provided between brackets. Tintinnid isolates are coloured according to their respective haplotypes based on ITS analysis. **(b)** Maximum likelihood rooted phylogenetic tree of the diatom ITS+28S-rDNA based on 854 aligned positions. Numbers at nodes are bootstrap values in percentage rounded up for 1000 bootstraps. The branch rooting the tree has been shortened for the more clarity and isolates are coloured according to their respective haplotypes.

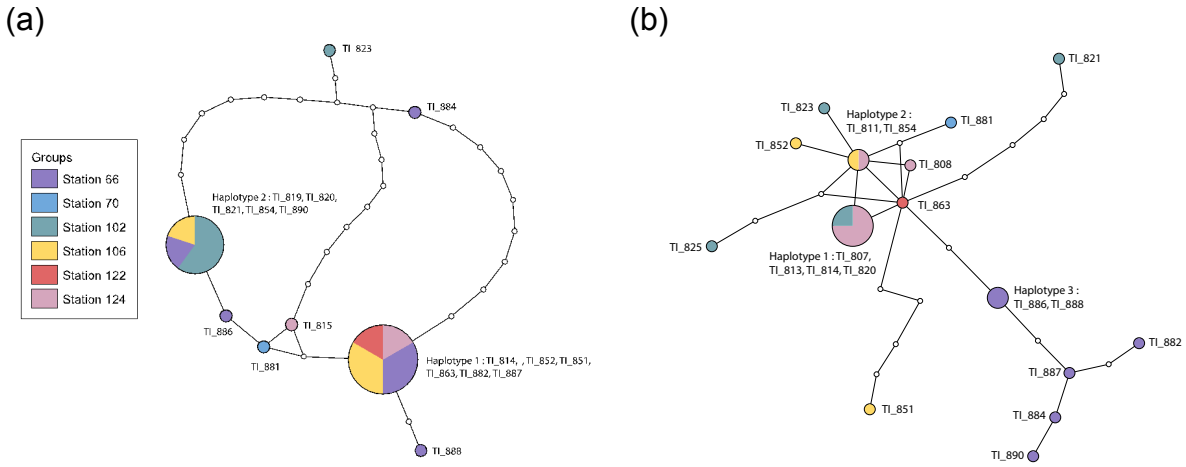


Figure 4.4. Statistical parsimony network with ITS+5.8S+28S rDNA.
The different haplotypes are labelled according to the corresponding isolated interactions and coloured by station of origin. (a) The tintinnid *Salpingella* sp. haplotypes display major clusters with Haplotypes 1&2 and small divergence overall (b) the diatom *Fragilariopsis doliolus* sequences were much more divergent, although several sequences from Station 124 clustered in one haplotype.

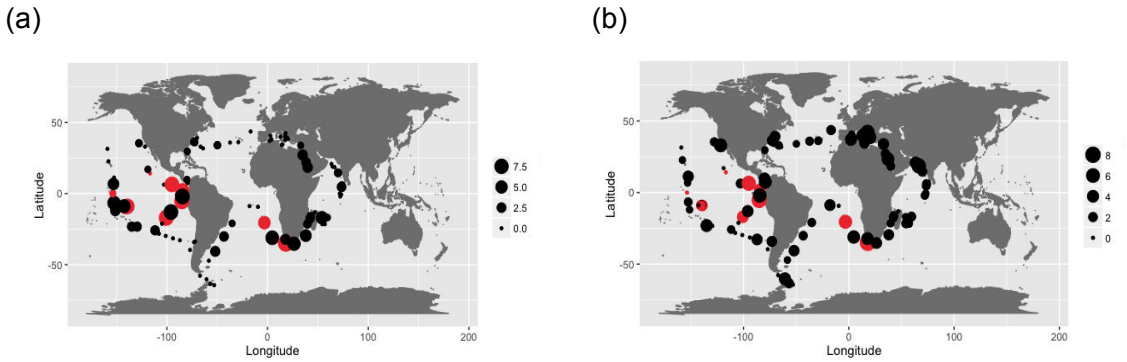


Figure 4.5. Spatial distribution of the isolated diatom and tintinnid metabarcodes across the 150 *Tara* Oceans stations.
Absolute abundance (fraction 20-180 micron in surface samples) was transformed according to the log(Abundance+1) formula with low to high abundance corresponding to small and big bubbles respectively. (a) Abundance of the diatom f2f8, assigned to *Raphid_pennate_X+sp* in *Tara* Oceans. Station 66 and target stations for further investigation based on co-occurrence of the two organisms in high proportion are coloured in red. (b) Abundance of the tintinnid a7cbc, assigned as *Choreotrichia_XX+sp* in *Tara* Oceans.

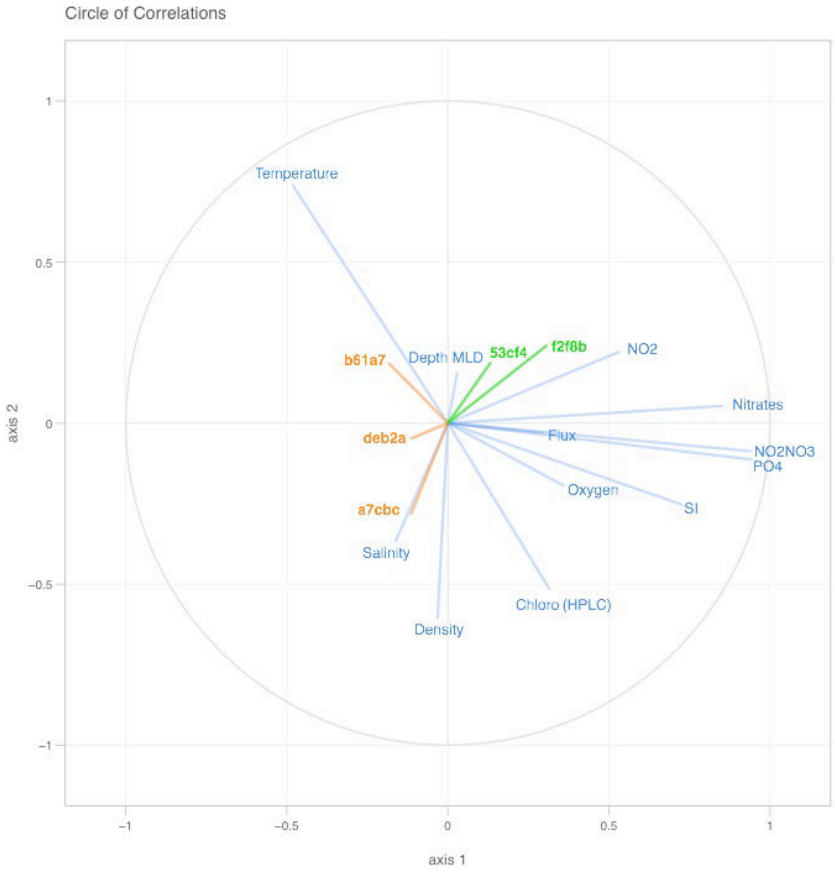


Figure 4.6. Circle of correlation for partial least square regression 2.
The five diatom and tintinnid barcodes were used as response matrix; tintinnid barcodes are labelled in orange and diatom barcodes labelled in green. The environmental parameters of surface samples in fraction 20-180 and 180-2000 microns were implemented as predictor variables and labelled in blue. Projection of explanatory variables on response variables reflect the sign and amplitude of the regression coefficients of responses on predictors. A7cbc is more abundant in high salinity samples, when the main abiotic driver for b61a7 is high temperature. Regression coefficients are available in **Table S7c**.

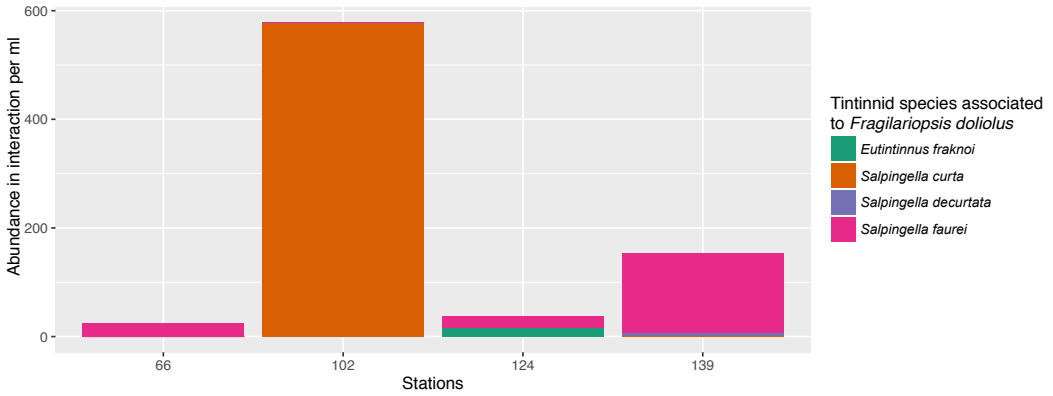


Figure 4.7. Number of diatom-tintinnid consortia across the ocean based on quantification of cell per mL.

Cells were counted from formol fixed net sample from 20-180 micron sized fraction, equivalent to 10L of filtered sea water. Four different species of tintinnids were counted when attached to *Fragilariopsis doliolus* across four different stations in South and North Pacific as well as Benguela current.

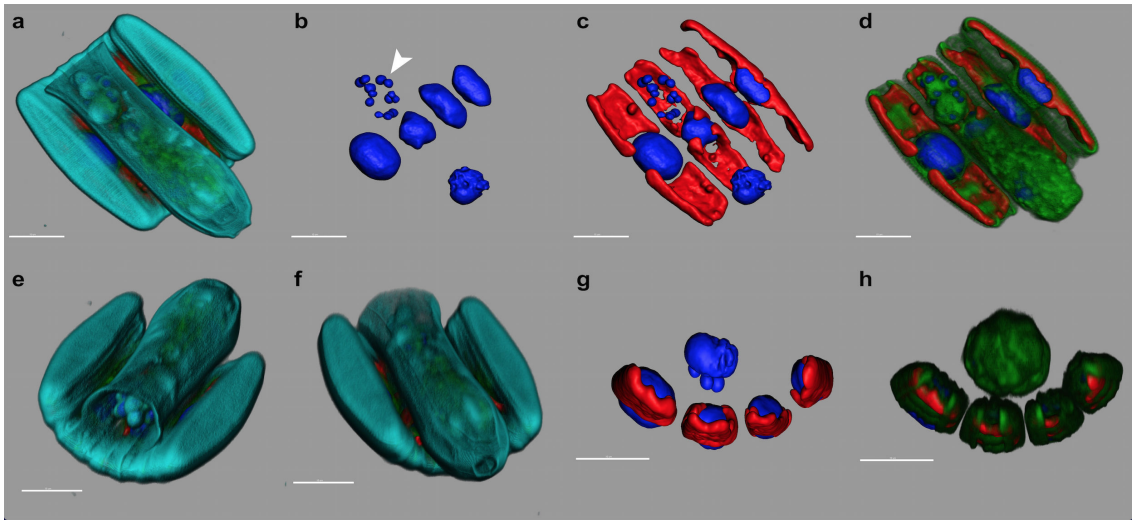


Figure S4.1. Microscopy images of a unique epibiotic assemblage between the diatom *Fragilariopsis doliolus* and the ciliate *Salpingella sp.* as initially observed in surface samples of Station 66 in Cape Agulhas.

The three dimensional (3D) reconstructions of the epibiosis were imaged with confocal laser scanning microscopy (CLSM). Figures (a-d) are taken under the same angle.(a) The diatom exhibits the *Fragilariopsis doliolus* barrel shape and the tintinnid *Salpingella* looking lorica. (b) DNA content from living cells revealed by Hoescht fluorescence signal, potential prey nuclei located in the tintinnid are indicated with a white arrox. (c) Chloroplasts of the diatom chain are highlighted by the red chlorophyll autofluorescence and displayed one on either side of the median transapical plane. (d) Membrane marker. (e-f) show both oral and aboral ends of the lorica. (g) Relative position of chloroplasts located on the convex side of the chain and (h) shows the interface between two organisms.

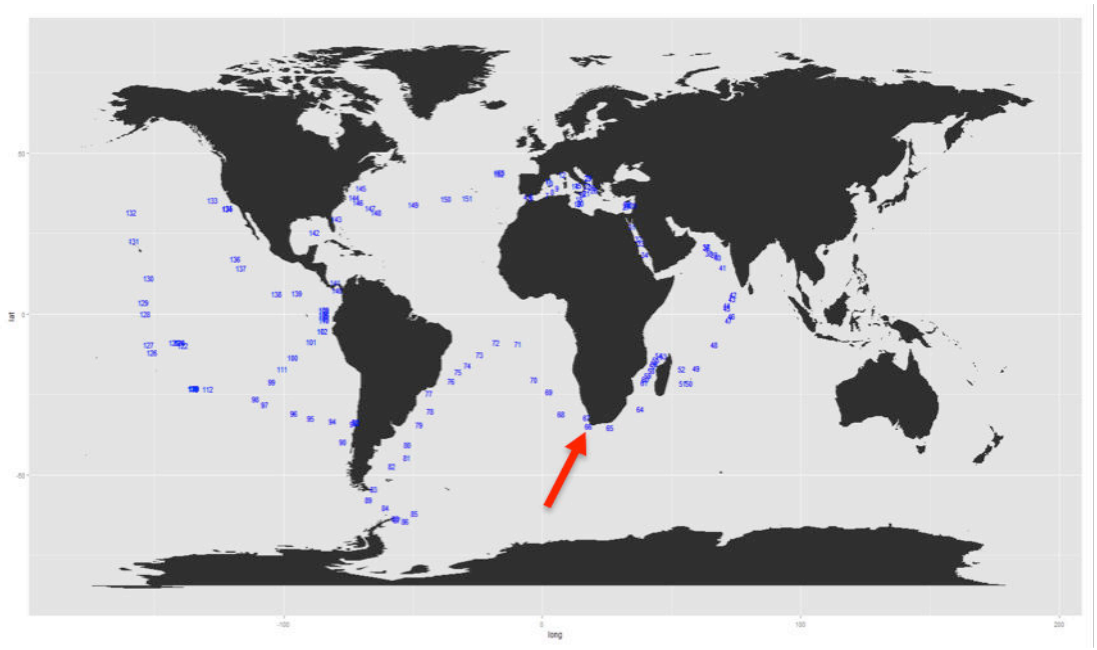
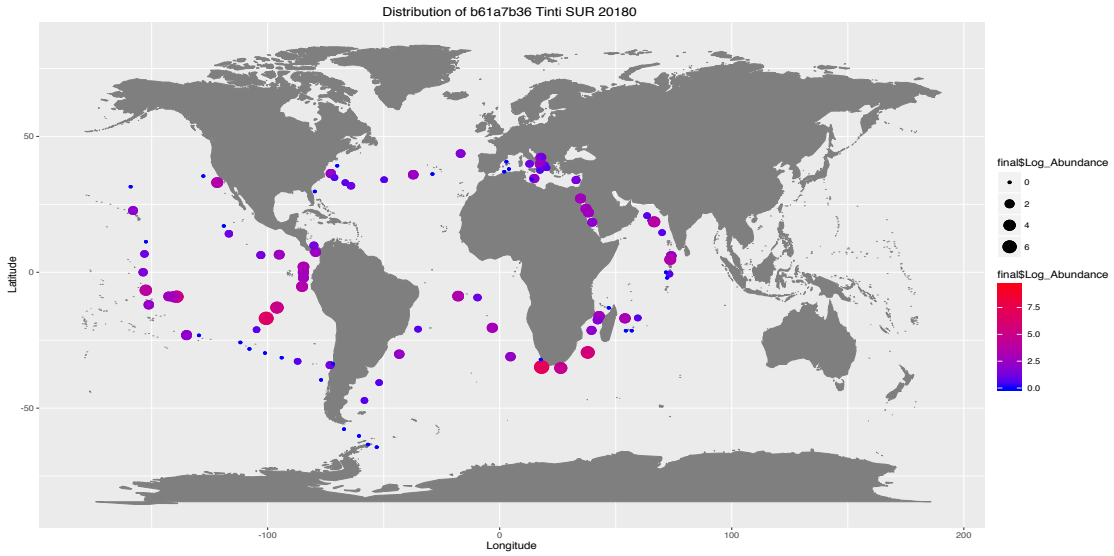
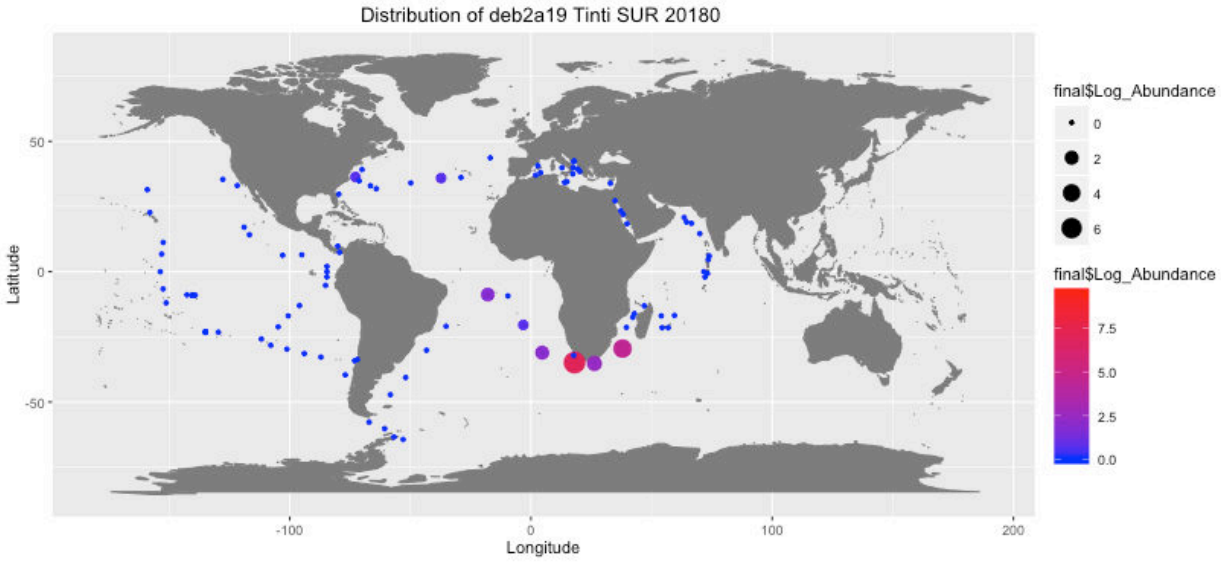


Figure S4.2. *Tara* Oceans 150 stations.
The red arrow indicates station 66 located in the Benguela upwelling current in which the diatom-tintinnid interaction was initially observed, in surface samples from the fraction 20-180 micron filter, preserved in formol-glutaralaldehyde.

(a)

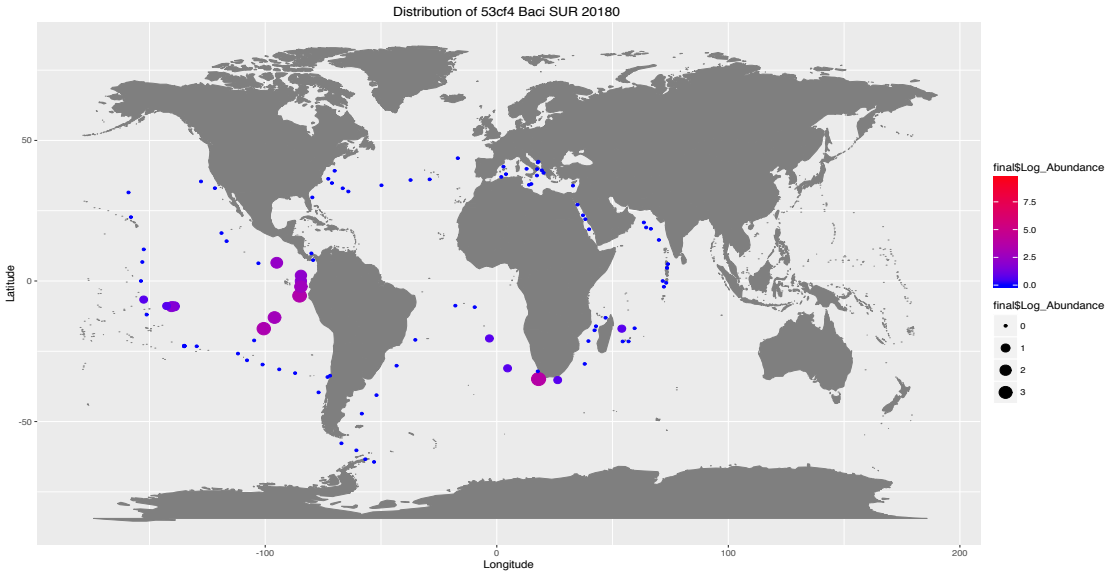


(b)



(see following page **Figure S4.3** for legend)

(c)



(d)

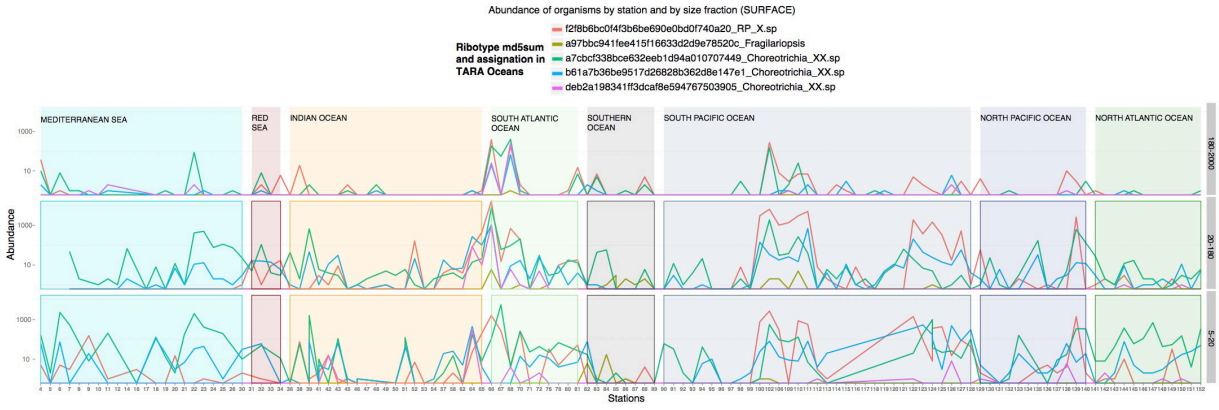


Figure S4.3. Spatial distribution of the V9 sequences obtained by single sequencing of the diatom – tintinnid consortia.

Abundance was obtained across the 150 *Tara* Oceans stations of fraction 20-180 micron in surface samples. Absolute abundance was transformed according to the log(Abundance+1) formula with low to high abundance corresponding to small and big bubbles respectively. (a-b) Abundance of the tintinnids b61a and deb2 respectivelent, both assigned to Choreotrichia_XX+sp in *Tara* Oceans. (c) Abundance of the diatom 53cf4, assigned as Raphid-pennate_X+sp. in *Tara* Oceans. (d) Merged view of the abundance of the five amplified barcodes, in three size fractions, across the *Tara* Oceans provinces.

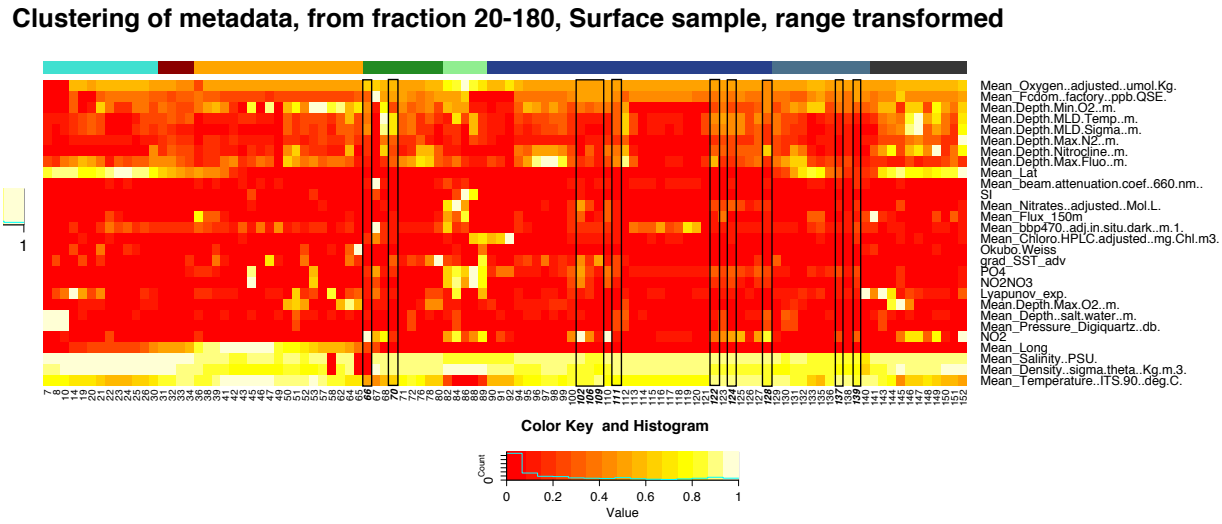
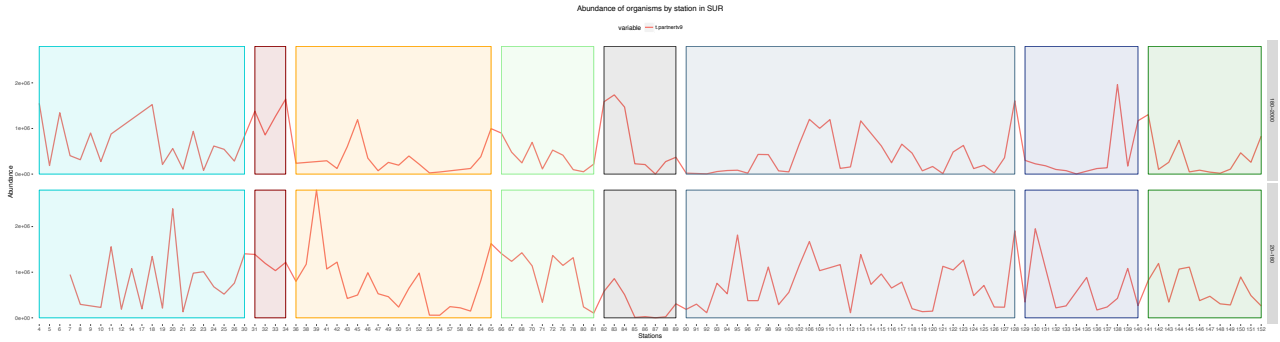


Figure S4.4. Heatmap of *Tara* Oceans 150 stations metadata in surface samples of fraction 20-180 micron.

Values of each environmental variable were standardized into range 0 to 1. Stations are indicated at the bottom abscissa. Colors at in the top axis correspond to the ocean province (from left to right : Mediterranean Sea, Red Sea, Indian Ocean, South Atlantic Ocean, Southern Ocean, South Pacific Ocean, North Pacific Ocean, North Atlantic Ocean). Stations in which the diatom-consortia was observed are framed in a black rectangle.

(a)



(b)

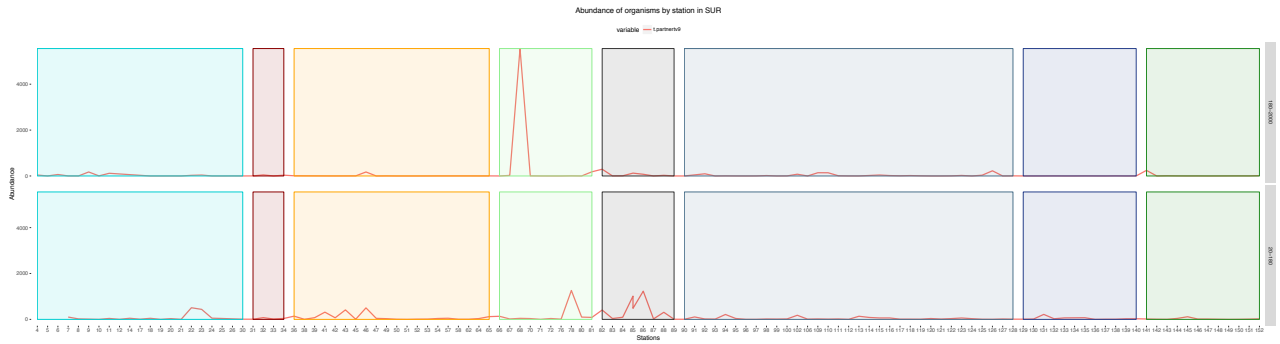


Figure S4.5. Spatial distribution of tintinnid predators and competitors in the *Tara* Oceans data.
(a) Absolute abundance of all Copepode barcodes merged together. (b) Abundance of all Oligotrichs barcodes grouped together. Absolute abundance in surface samples of both fractions 20-180 (lower panel) and 180-2000 microns (upper panel). The colours correspond to ocean provinces, as in **Figure S4.3(d)**.

.

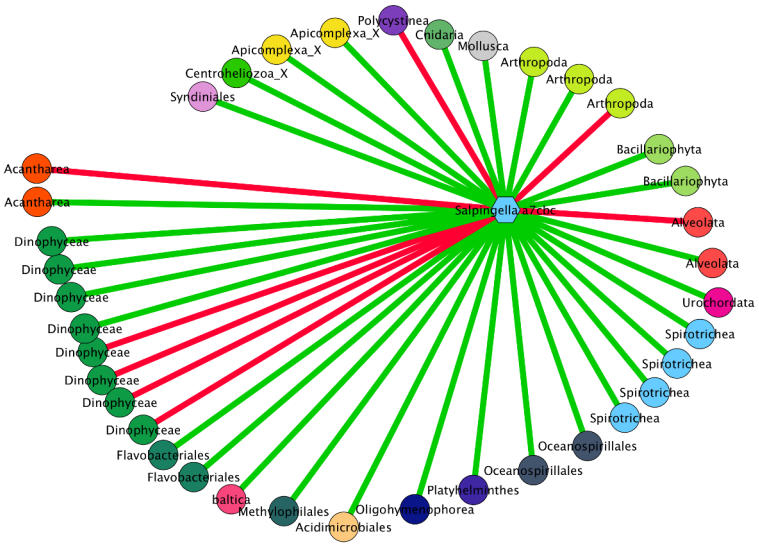


Figure S4.6. A7cbc co-occurrence network.
Co-occurrence network extracted from the *Tara* Oceans interactome (Lima-Mendez et al., 2015). Each node represents a unique barcodes and is coloured and named by its taxonomic group. Red edges represent mutual exclusions, and green edges represent highly positively correlated barcodes.

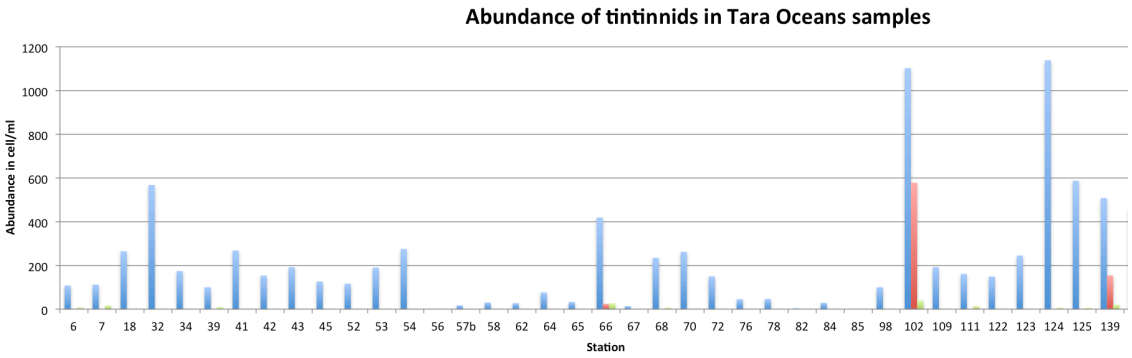


Figure S4.7. Tintinnid cell counts in *Tara* Oceans stations.
The counts are based on quantification per ml of formol fixed samples in 20-180 micron size fraction, equivalent to 10L of filtered sea water. Total tintinnid counts in blue, diatom associated *Salpingella* species in red and single *Salpingella* species in green.

Chapter 5: Single cell genomics to explore diatom biotic interactions

Summary

Abstract

Microbial community genomics have opened the way for integrated studies that connect multiple scales of biological information, from metabolic pathways to ecosystem modeling. However, understanding biotic interactions at the single-cell level is still in its infancy, thus limiting the successful integration of information from the cell to the ecosystem. The study of single biotic interactions in natural samples faces several challenges, amongst which the identification of partners of interactions, the metabolic processes involved and the importance of host associated bacterial communities. By adopting high throughput sequencing of individualized interactions isolated from environmental samples fixed in formol-glutaraldehyde and ethanol, I propose a novel approach to identify both eukaryotic and prokaryotic entities forming biotic associations, based on three different types of diatom associations observed in *Tara* Oceans samples. I analyze the full transcriptome of 147-pooled diatom interactions presented in Chapter 4 and provide preliminary results on the functional content of such a consortia.

5.1. Introduction

Microbial community genomics have opened the way for integrated studies, that connect multiple scales of biological information, from metabolic pathways to ecosystem modeling (DeLong, 2005) and have rapidly evolved in the past decades thanks to high throughput sequencing and computational advances (Raes et al., 2008). Many of these holistic approaches in plankton biology rely upon population level frameworks. However, understanding biotic interactions at the single-cell level is still in its infancy, which limits the successful integration of information spanning the cell to the ecosystem (Brehm-Stecher et al., 2004). Biotic interactions in the plankton can be considered as small systems, so much so that comprehensive approaches ought to be developed for a better understanding of their impact on organismal evolution and function. The study of single biotic interactions in natural samples faces several challenges, amongst which the identification of partners of interactions, the metabolic processes involved, and the importance of host associated bacterial communities (Hilton et al., 2013; Carotenuto et al., 2014).

Accurate identification of species is critical for plankton biogeography studies, in particular regarding partners of biotic interactions. For more than a century, identification was ensured by recognition of morphological traits, performed by an expert – but rare - taxonomist in best of cases, who's knowledge is both crucial and troublesome to transfer. On the other hand, the arrival of DNA-based barcoding methods has transformed identification of plankton in a rapid, inexpensive highly resolved task, though it is not without problems either (see Annex A). The discovery of cryptic species – genetic diversity hidden behind apparent morphological homogeneity in one or a group of species – and planktonic species difficult to identify by eye, has put forward the necessity of combining simultaneous morphological and molecular methods for plankton identification (McManus et al., 2009). This transition from morphological to molecular techniques is accompanied by major challenges in terms of DNA extraction, particularly from formalin-diluted preserved samples, a common fixator used in plankton studies also known to degrade DNA (Bucklin et al., 2004). This brings forward the need to optimize new approaches in order to rapidly and reliably identify partners of *in situ* isolated interactions, by keeping track of both the morphology and the molecular identification of the partners.

Moreover, if we consider the biotic interaction system as a whole, interest regarding host-associated bacterial communities arises. Plant microbiomes are critical to host adaptations, and influence plant health as well as productivity (Compant, 2010; Berendsen, 2012). In phytoplankton, host-associated bacteria vary between strains that display different functional traits (Kaczmarska et al., 2005). It is likely that biotic interactions between two major partners benefit as much from the functions carried out by the partner organisms than from the bacteria associated with them. If, and how, the interaction-associated bacteria vary through space is also hardly explored, as it relies upon large-scale sampling initiatives.

Beyond phylogenetic identification of major partners and bacteria, understanding the metabolic processes engaged in biotic interactions is of key interest, in order to bridge the gap between functional traits and community ecology. Functional genomics have enabled

the era of "transcriptomics", that is the sequencing of a targeted or complete set of expressed genes that ultimately track the transcriptional activity of an organism (Wang et al., 2009). Current studies of diatom biotic interactions monitor gene expression variation based on co-cultures exposed to a range of changing parameters (Carotenuto et al., 2014). But if we ultimately seek to understand how biotic interactions regulate the fate and activity of marine microbes, it is timely to investigate gene expression of microbial associations in the natural environment.

By adopting high throughput sequencing of individualized interactions isolated from environmental samples fixed in formol-glutaraldehyde and ethanol, I propose a novel approach to identify both eukaryotic and prokaryotic entities forming a biotic association, based on three different types of diatom associations observed in *Tara* Oceans samples. I further analyze the full transcriptome of 147 pooled diatom interactions presented in Chapter 4 and provide preliminary results on the functional content of such a consortia.

5.2. Materials and methods

5.2.1. Isolation of single diatom interactions and DNA extraction

Based on micropipette manipulation, three different diatom interactions were isolated from Tara Oceans samples fixed in both formol-glutaraldehyde and ethanol. A total of 14 interactions were isolated from formol-glutaraldehyde samples and 40 from ethanol samples (**Table S1**). DNA was extracted using the MasterPure™ Complete DNA and RNA Purification Kit protocol (Epicenter) adapted to formol and ethanol fixation, respectively.

5.1.1. HTS of unidentified eukaryotic partners and host-associated bacteria

A first batch of interactions isolated from formol-glutaraldehyde preserved samples was PCR amplified with 18S-V9 primers (1389F - 1510R, fragment length 170) used in the Tara Oceans expedition (de Vargas et al., 2015) and sent for high throughput sequencing with duplicate PCR (**Table S1**), using genomic DNA of the Phaeodactylum tricornutum Pt1 8.6

(CCMP2561) culture as a positive control. Pt1 8.6 is a clonal, axenic culture of a single species of diatom so we expected no traces of contaminants in the results; however, multiple 18S copies are expected. Collaborators in Station Biologique de Roscoff performed sequencing and annotation of this batch.

The second batch of interactions isolated from ethanol preserved samples was amplified with 18S-V9 primers (1389F - 1510R length 170) as well as 18S-V4 primers (TAReukF1-TAReukR length 380 (Massana et al., 2015)) and triplicates of PCR products were used for sequencing (**Table S1**). The V9 region was amplified in order to identify the interacting partners in the Tara Oceans metabarcoding dataset, whereas the V4 region was sequenced to improve phylogenetic assignment. Technical replicates were performed to enable rigorous interpretation of amplicon variants (Decelle et al., 2014). DNA extracts from the same ethanol fixed interactions were used to identify the prokaryotic community by amplifying the 16S-rDNA using the Fuhrman Primers (515F-926R length 411 (Parada et al., 2015)). Sequencing was performed using dual indexed primers on miSEQ Illumina platforms (Protocol, J.Raes Lab).

5.1.2. RNA-Seq of diatoms associated to heterotrophic ciliates

Sequencing and assembly. A total of 147 individual diatom-tintinnid associations were isolated in the microscope using micropipettes from ethanol-preserved samples. DNA and RNA were simultaneously extracted from the pooled collection of cells using a combined kit (J. Poulain, Génoscope). After checking that RNA was not contaminated with DNA, cDNA synthesis was performed using the Kit Ovation® RNA-Seq System V2 as concentration of total RNA was approximately 400 picograms, following which an Illumina library was built. Sequencing was performed in two different runs (1) using 2% of a HiSeq 2500 to validate the Illumina library, and then (2) using 25% of a HiSeq 2500 rapid run in 2*100bp for a higher coverage. Assembly of both sequencing reactions (2% and 25% of rapid run) was performed using Velvet 1.2.07 (Zerbino et al., 2008) - kmer of length 63, insert size 200 - and Oases 0.2.08 (Schulz et al., 2012) to treat transcript isoforms. A second assembly was done using the Trinity Software (Haas et al., 2013) as the algorithm used for de novo assembly can highly impact the remapping rate of reads on assembled contigs

Taxonomic and functional annotation. Reads assembled with Velvet/Oases were cleaned from ribosomal sequences and taxonomically annotated using the metatranscriptome pipeline developed by Genoscope to assign *Tara* Oceans metatranscriptomic data (E. Pelletier). Functional annotation was performed using the InterProScan software (Jones et al., 2014) after transcripts were translated using Transdecoder (Haas et al., 2013) that detects different Open Reading Frames (ORFs). InterProScan uses InterPro (Mitchell et al., 2014), the most up to date database that integrates many sources of functional annotation (Pfam, Gene3D, Panther, ProSite) giving access to protein families, as well as domains that a protein contains. Query sequences were compared with all the members of the database: the higher the consensus of functional annotation amongst different sources for a given protein, the more reliable it could be considered.

5.3. Preliminary results

5.3.1. Abundant and unidentified interactions in the *Tara* Oceans environmental samples

One of the analytical platforms of the *Tara* Oceans project involves the high-throughput analysis by confocal laser scanning microscopy of formol-glutaraldehyde-fixed fractions enriched in nano- and micro-plankton. The screening of the samples led to the observations of several diatom interactions of ecological interest. One of them was observed in *Tara* Oceans Polar Circle samples in the North East of the Kara Sea (Station 188, Lat 78.304 / Long 91.725) and involved a chain-forming diatom (probably a pennate diatom) that showed the presence of potential parasites in high abundance (**Figure 5.1 A1-C2**) under epi-fluorescence microscopy. In this case, we cannot completely rule out the possibility that such cells may be a consequence of sexual reproduction, although auxospore formation is believed to be restricted to centric diatoms. A second interaction was observed in the oligotrophic Station 72 in the Southern Atlantic Ocean and involved the diatom genus *Chaetoceros* in association with small epiphytes, from which we could distinguish two major types: one in which the epiphytes were attached to the setae (**Figure 5.1 D1-D3**), and another in which they were directly attached to the diatom frustule (**Figure 5.1 E1-E3**).

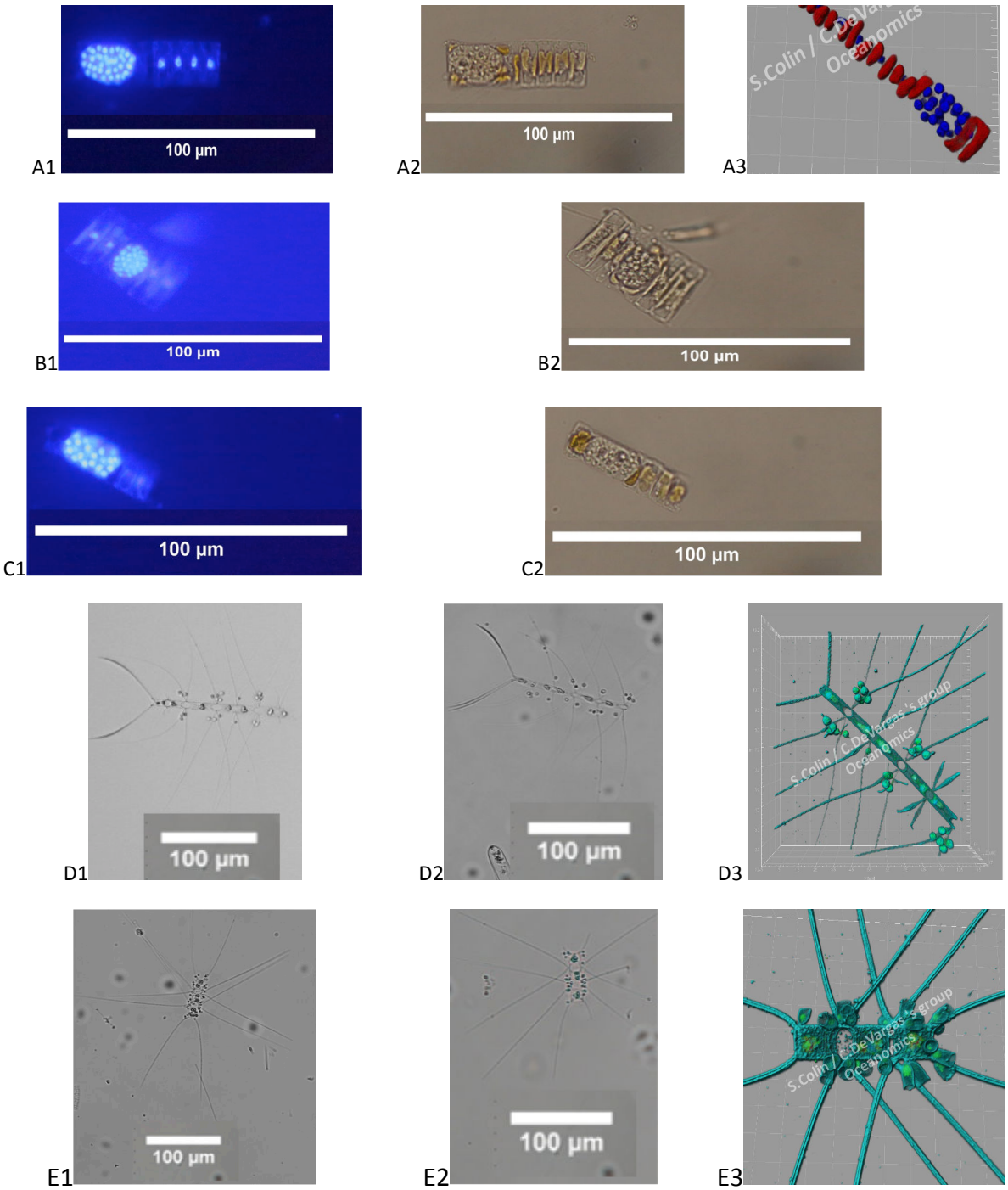


Figure 5.1. Interactions isolated from formol-glutaraldehyde samples.
A1-C2: Interaction involving diatoms and potential parasites in the Arctic. D1-D3: interaction between *Chaetoceros* sp. and epiphytes attached directly to the setae. E1-E3: interaction between *Chaetoceros* sp. and epiphytes attached to the frustule. A3-D3-E3: Confocal Laser Scanning Microscope images with nuclei (blue), cellular membranes (green), chlorophyll (red), courtesy S.Colin. A1-B1-C1: Epifluorescence Microscope images, nuclei (blue).

5.3.2. HTS enables phylogenetic identification of partners in direct biotic interactions

Amplicons obtained from high throughput sequencing of formol-glutaraldehyde fixed samples are shown in the **Table S2**. Amplicons were analyzed by BLAST against the *Tara* Oceans metabarcoding database (lineaget) and against NCBI (lineageb). The analysis was done on all the reads, but the table only gives barcodes that have above 100 reads across the interaction replicates.

The positive controls (F15-F16) contained genomic DNA extracted from *Phaeodactylum tricornutum* Pt1 8.6 (CCMP2561) lowered to a concentration of 10 nanograms /μL before 18S-V9 amplification (preserved in water). Out of the four replicates, 391 different barcodes were assigned to CCMP2561 with one highly abundant representative barcode followed by a very skewed rank abundance curve (**Figure 5.2**).

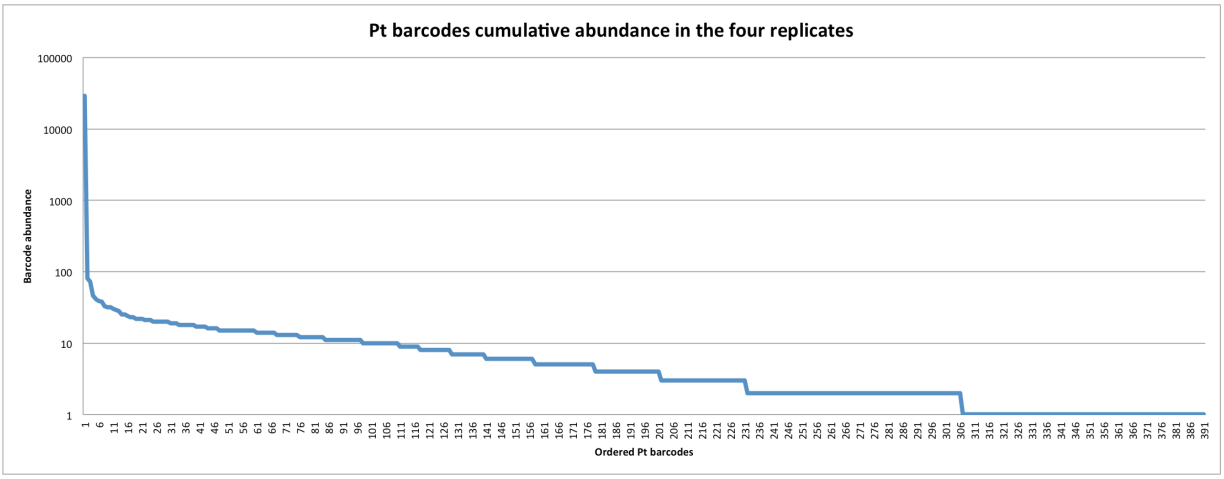


Figure 5.2. Rank abundance of *Phaeodactylum* barcodes obtained by high throughput sequencing.

A phylogenetic network estimation using statistical parsimony was built based on the 391 aligned *Phaeodactylum* sequences using the TCS software (Posada et al., 2005) to see how distant were the different sequences from each other (**Figure 5.3**). Very distant sequences (6 bp from the representative barcode) contain large deletions in the alignments. Overall,

each sequence had two to three base pair differences with all the other variants, resulting in this highly clustered network.

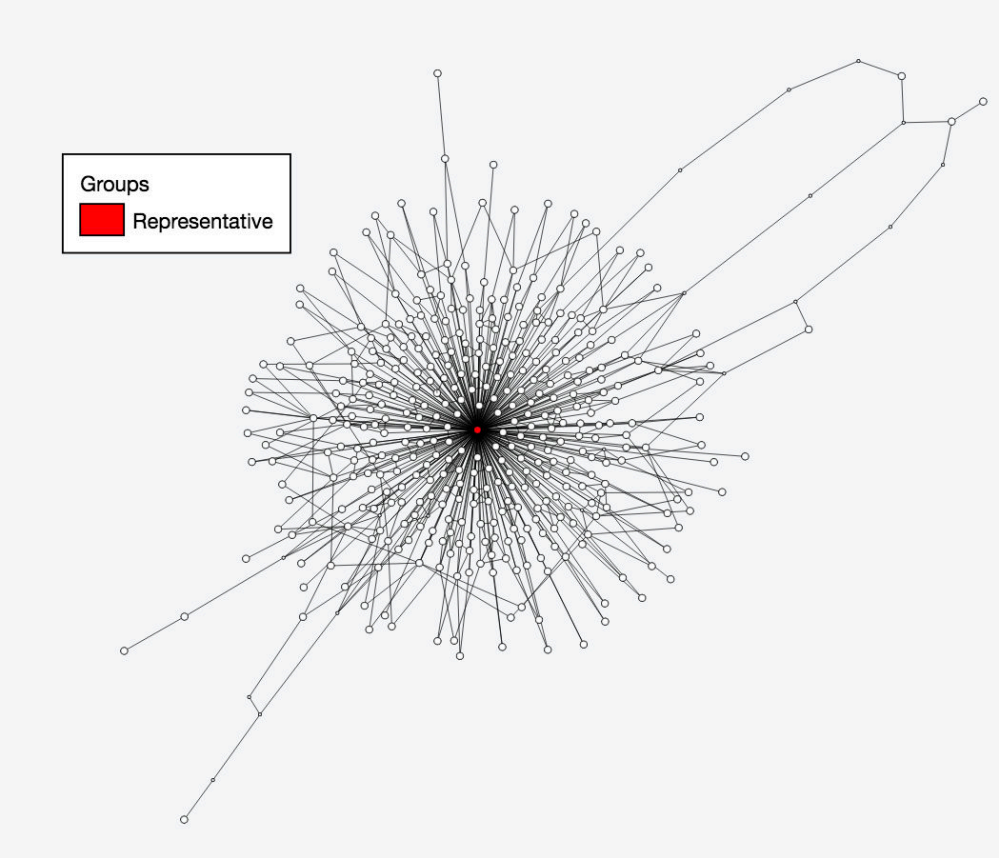


Figure 5.3. Network of *Phaeodactylum* variants using statistical parsimony.
The most abundant and representative barcodes is colored in red.

The number of reads for the most representative *Phaeodactylum* sp barcode (bd6a2410e9b77fc2dba49500ecafe8a0) was highly variable between two replicates, as illustrated by F15 replicates F15N1 and F15N2, containing 15,243 and 3,693 reads, respectively, or between F16N1 and F16N2, containing 0 and 10,365 reads, respectively (**Table S2, Control**). Sequences of *Mus musculus* and *Uncultured fungus*, considered as contaminants, are present in low abundances with orders of magnitude in read counts 5 to 200 times lower than the representative diatom barcode.

For the Diatom - Tintinnid interaction (**Table S2, Diatom-Tinti**), reads assigned to Fungi, Uncultured Fungus, Metazoan, Bacteria, were filtered out. Sequences of Uncultured fungus

clones referenced in NCBI came from indoor environments and *Malassezia* is a genus of fungi naturally found on the skin surface of many animals including humans. Successive filtering steps revealed the barcodes of interest a7cbc (Tintinnid) and f2f8 (Fragilariopsis) to both be present in the *Tara* Oceans data (**Table S2, Diatom-Tinti, TIF2_N1, TIF2_N2, TIF5_N1**). Contrary to the positive control of Pt1 8.6, the read counts of contaminant sequences were at least two orders of magnitude more abundant than the diatom and tintinnid barcodes. Replicates for TIF2 contained 29 and 93 tintinnid reads as well as 45 and 77 diatom V9 reads. In total, twenty amplicons were assigned to the target *Raphid-Pennate* with the GenBank Accession number gb|JQ782062.1 and a cumulative abundance of 160 reads across the ten replicates. Ten amplicons were assigned to the target tintinnid, with the GenBank Accession number gb|KF662528.1 and a total cumulative abundance of 154 reads across the ten replicates.

For the Diatom - Parasite interaction,(refer to **Figure 5.1 A1-C2**) we cannot filter out Fungi as they are strong parasitic candidates. The most abundant diatom barcode (cf7d6062d57f5919cac2af3d59d30b27) was assigned to *Araphid-Pennate*, and the GenBank ID gb|KC771163.1 but also has similar BLAST scores to the diatom *Fragilaria* sp.. Four amplicons have this assignation, for a total abundance of 275 reads across 10 replicates. I then restricted the analysis to barcodes that co-occurred in the same replicates; a potential candidate was assigned to the Ascomycota phylum of the Fungi kingdom (ff30009a2ecee7eaf0e2e6c2d48c4192), most of which are parasitic, and some of which have adapted to arctic environments (Zhang et al., 2015). The total abundance across 10 replicates was of 247 reads. The same filtering strategy was adopted for the Diatom - Epiphyte interaction (refer to **Figure 5.1 D1-E3**) with no distinction of both morphotypes. The absence of a clear conserved phylogenetic signature across a majority of replicates complicates interpretation of the results. Two amplicons were assigned as *Chaetoceros rostratus*, for a total cumulative abundance of 11 reads across 8 replicates. However, I could not determine the true nature of the epiphytic partners. Sequencing results of ethanol fixed samples for the 18S-V9, 18S-V4 and 16S are not available yet but will certainly confirm or refute the above preliminary candidates and filtering method with less ambiguity.

5.3.3. RNA-Seq of diatoms associated with heterotrophic ciliates

The transcriptome of 147 diatom-tintinnid associations was sequenced, leading to a total of 135 million reads (**Table 5.1**). As no mRNA isolation nor rRNA depletion was performed prior to library creation, rRNA that compose a large majority of RNA molecules present in a cell are highly abundant (111 million reads).

Type of transcripts	N Reads	Velvet/Oases Assembly			Trinity Assembly		
		Assembly %	N Contigs	Mean Size of Contigs	Assembly %	N Contigs	Mean Size of Contigs
Non rRNA Transcripts	24 million	83%	16498	646	91.3%	32996	531
rRNA Transcripts	111 million	93%	7507	423	NA		

Table 5.1. Summary of sequencing information for 147 pooled diatom-tintinnid interactions.

Taxonomic affiliation of rRNA-free transcripts using the *Tara* Oceans pipeline revealed that over 60% of the transcripts are unassigned, i.e., they do not have a single match in the reference database that enables their taxonomic affiliation. 9% of the transcripts match bacteria. The remaining 31% of transcripts match eukaryotic sequences, representing 5,159 contigs (**Figure 5.4**). Amongst these, 31% of eukaryotic matches, 30% are unassigned eukaryotes, 20% are fungi, metazoans and dinoflagellates, 23% are assigned to diatoms and 24% to ciliates. After *in silico* transcription and translation, the 16,498 contigs were translated in 2,674 unique proteins for which the available GO terms are summarized below.

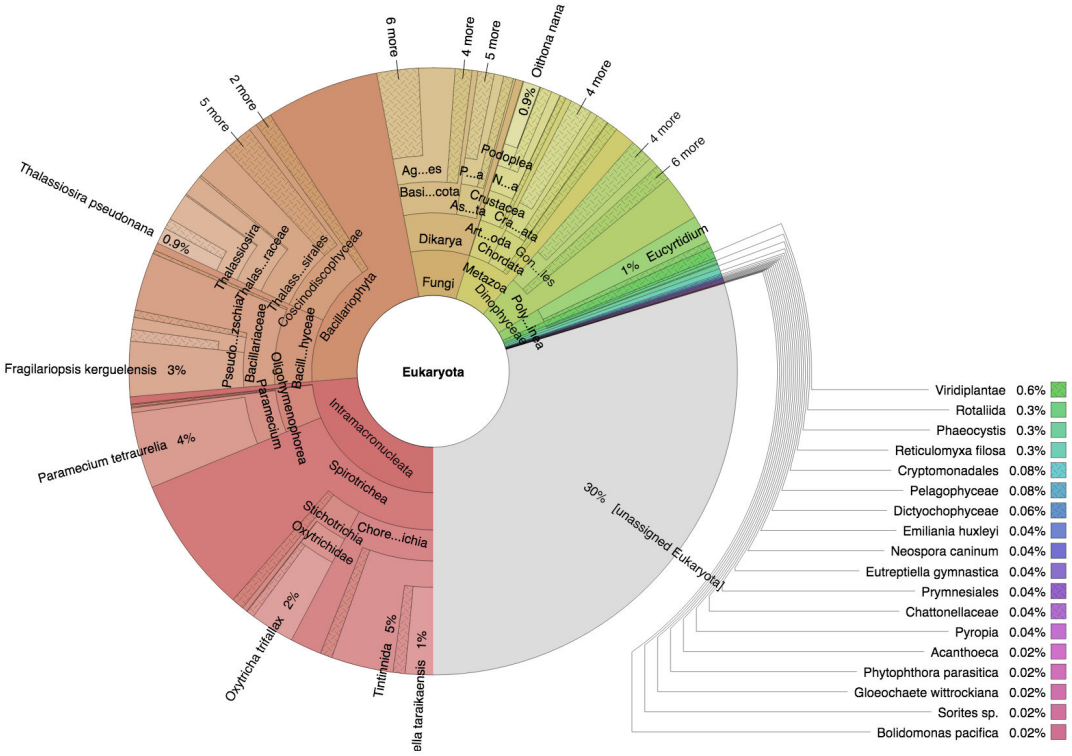


Figure 5.4. Taxonomic affiliation of non rRNA eukaryotic transcripts of 147 diatom-tintinnid pooled transcriptomes.

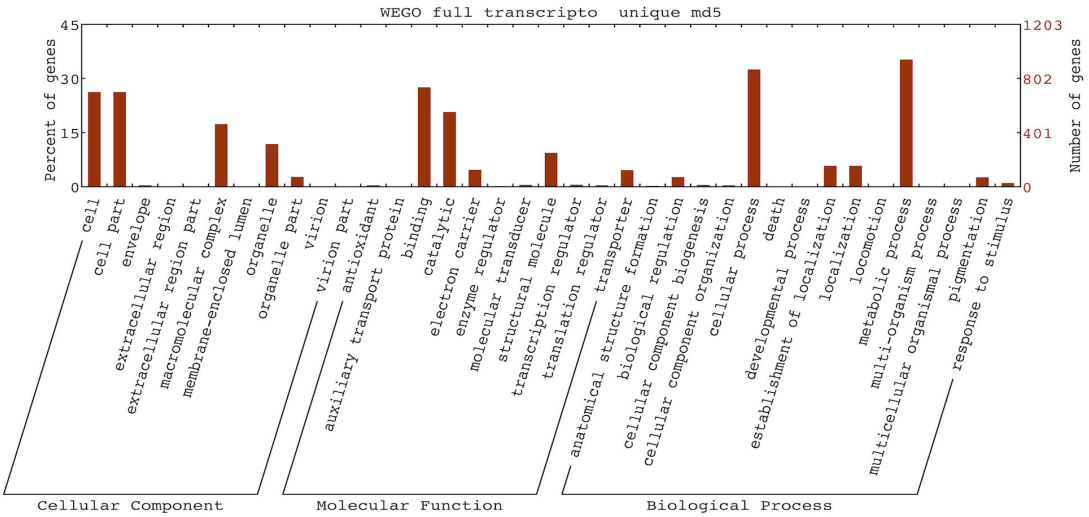


Figure 5.5. High level GO terms for three major ontology type (level 2 GO).
Transcripts corresponding to unique proteins were kept to avoid redundancy from InterPro output. The transcripts were assembled with Velvet/Oases.

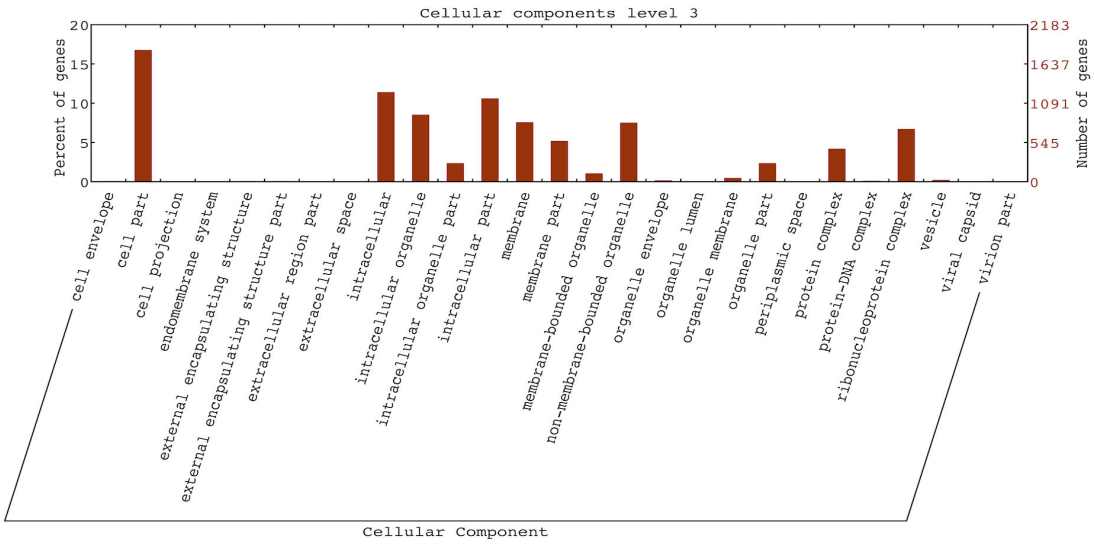


Figure 5.6. Details of GO from cellular components subtype (level 3 GO).

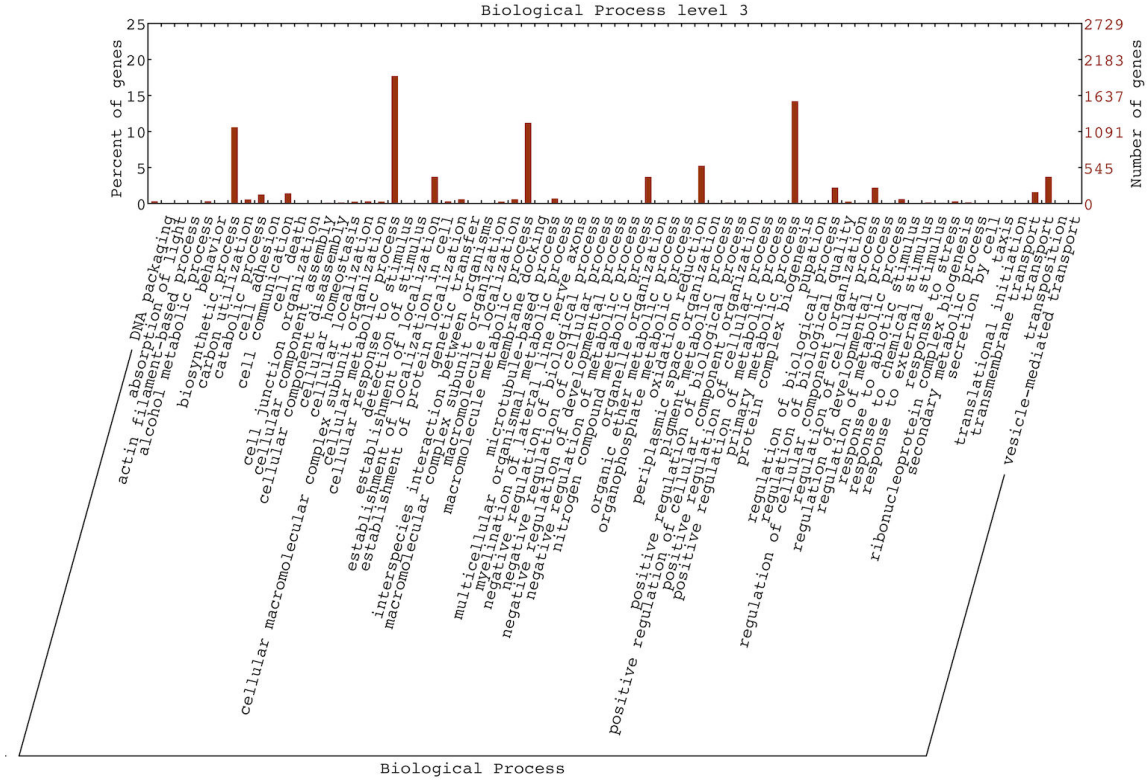


Figure 5.7. Details of GO from biological processes subtype (level 3 GO).

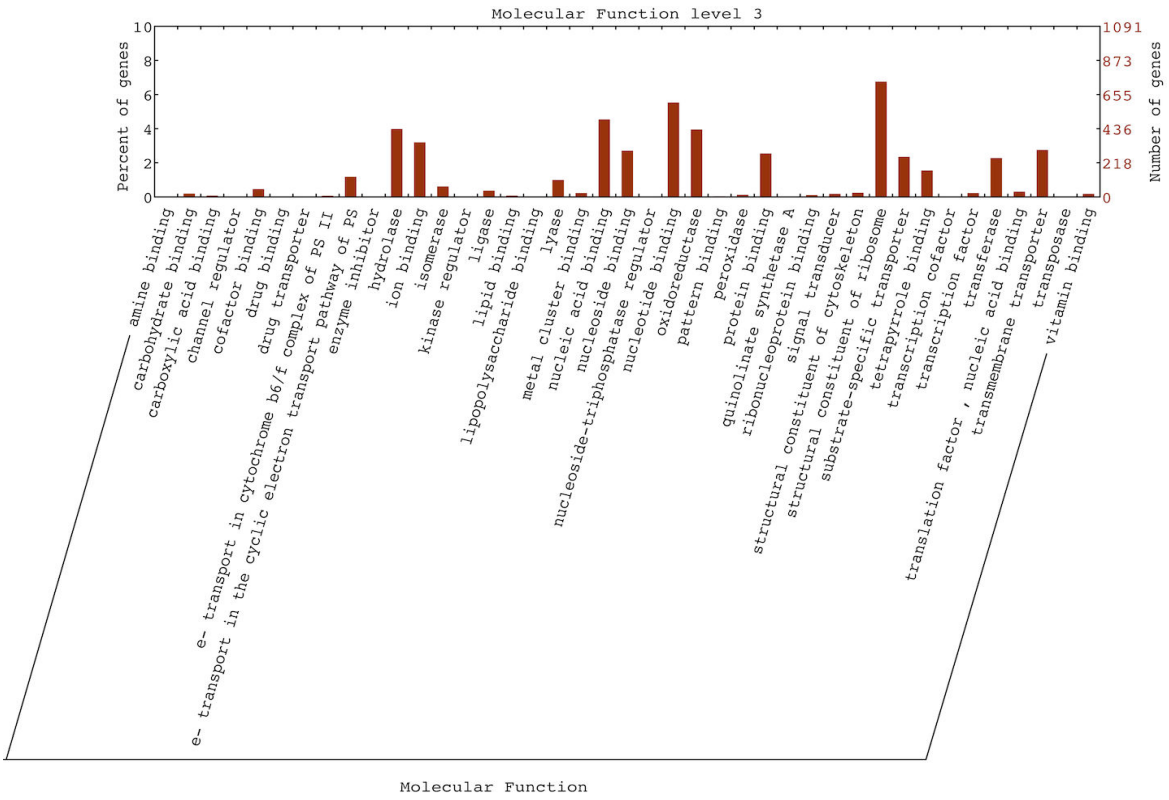


Figure 5.8. Details of GO from molecular functions subtype (level 3 GO).

The analysis of RNA-Seq at high GO levels is consistent with other transcriptomic data from diatom cultures with a high number of transcripts related to Cell part, Organelle, Binding, Catalytic functions, Cell processes, Metabolic Processes. Differences arise with an increased prevalence of macromolecular complexes (PSI, PSII, Proton Transport) and Increased structural proteins (ribosome). 11 transcripts related to lectins proteins were found and one in particular to C-Type lectins. C-Type lectins are carbohydrate-binding proteins involved in a range of functions such as cell adhesion, pathogen recognition and known to be involved in symbiotic interactions in Metazoa (Meyer et Weis, 2012) and were recently found in transcriptomic studies of Rhizaria symbiotic taxa (Balzano et al., 2015). Many molecular functions related to iron ion binding (GO:0005506) and zinc ion binding (GO:0008270) are found, as well as inorganic cation transmembrane transporter (GO:002290). However, a more refined analysis of diatom specific and tintinnid transcripts is necessary to make reliable conclusions.

5.4. Discussion

5.4.1. Species identification despite variants and contaminants

High throughput sequencing (HTS) of diatom biotic interactions amplified with universal primers based on formol-glutaraldehyde samples revealed many contaminant sequences with abundances orders of magnitude higher than expected. Due to the omnipresence of fungal sequences in many other PCR amplified samples fixed in formol-glutaraldehyde (personal communication) it is likely that the initial fixator was the source of contamination, rather than the downstream isolation and DNA extraction for which several rinsing and control of PCR reactants were performed. However, this does not exclude the fact that these steps add additional contamination. Indeed, the positive control using *Phaeodactylum tricornutum* diatom should show no traces of other DNA except the diatom, yet small amounts of fungal sequences were also present. This positive control was a culture extract, from several cells of a "single" population. Many factors can cause the appearance of the 391 V9 variants: cultures can display high intra-individual variability and multiple 18S copies per individual (*Phaedactylum* has 3-4 18S copies, personal com.); PCR and sequencing errors can occur. Attributing the different variants to each of these processes has been attempted (Kennedy et al., 2014) but I would expect all these variants except the most abundant ones to be biases rather than to represent true diversity.

The HTS procedure was successful in recovering barcodes of the partners in the case of the Diatom - Tintinnid consortia. However, they would have been hard to discover without prior knowledge of the sequences, as each replicate contains fungal sequences representing 100 times more reads than the barcodes of interest. It should be noted that these interactions were isolated with manual micropipette, and not the micromanipulator on the inverted microscope (described in Chapter 4), which also adds uncertainties in the cleanliness of the isolation. Future results of isolated interactions from ethanol samples, with the micromanipulator, should be less prone to contamination.

In summary, high throughput sequencing from formol-glutaraldehyde samples can provide taxonomic assignment of individualized interactions, when data is treated with a set of

filtering steps based on several decision criteria such as: the plausibility of observing certain reads, prior idea of which taxonomic group the partners belong to, abundance of specific reads, co-occurrence with another identified partner. Ultimately, optimizing this procedure will be necessary for the diatom-parasite interactions, consortia for which fluorescence observation is mandatory (and thus formol-glutaraldehyde fixation) to identify potentially infected cells. I was able to isolate them without fluorescence, yet bright-field pictures are insufficient to distinguish the parasites properly. For other types of eukaryotic interactions, such as diatom–tintinnids, high throughput sequencing from ethanol-fixed samples is likely to be much more effective than cloning, as contamination will certainly be lower and plausible reads easier to detect (unlike parasites or nanoflagellates). Continuous progress in the field of sequencing creates the opportunity to move forward and consider whole genome sequencing of both partners, asking whether this peculiar association leaves imprints in the genome of one of the partners, much like recent exciting discoveries in diatom-diazotroph symbiosis (Hilton et al., 2013).

5.4.2. RNA Seq of diatom-tintinnid interactions

The initial idea behind the pooling of different diatom-tintinnid associations was to increase the RNA material. Of course, many bacteria are also likely to be attached to the diatom biofilm, and many nanoflagellates and dinoflagellates can be consumed by the tintinnid. It is therefore expected to find moderate to high eukaryotic and prokaryotic (no oligo-dT selection) diversity in the phylogenetic annotation of the transcripts, as indeed reported. The initial concentration of RNA present in the sample was barely detectable by Bioanalyzer, around 400 pg total RNA; therefore assessing the quality of RNA was not possible. Yet, we can consider that the identified transcripts represent the functional pool of the association.

To refine the biological interpretation of the transcripts, I plan to map the sequences against both *Fragilariopsis* and *Paramecium* transcriptomic sequences (MMETSP; Keeling, 2014), and to assign contigs at a finer taxonomic and functional level, subsequently decreasing the proportion of unknown contigs. In particular, I will investigate over- or under-expressed genes with respect to culture conditions of *Fragilariopsis* cultures. This should reveal additional features with respect to metabolic and functional genes in the consortia. Investigating transcripts encoding glycoproteins of the extracellular matrix and involved in

cell adhesion such as fibronectin, fascin, fasciclin will be of interest; molecular functions such as nutrient and metabolite transport should also be explored. Ultimately, even though this is still extremely challenging, single interaction transcriptomics from ethanol preserved samples could be an extremely exciting future step to apply on the *Tara* Oceans data (Kanter et al., 2015), and has already been applied on *Emiliana huxleyi* cultures infected by viruses (Assaf Vardi, personal com). By single cell sequencing of the diatom, the tintinnid, and the individualized consortia, we are likely to gain insight on what the association really brings in terms of cost and benefit. Metabolomics on this association could have brought insight about exchanged metabolites, likewise NanoSIMS analysis from *in situ* live samples would help understand the microbial activity of both partners, however our case study is likely to be constrained by the non-culturability of both species and the difficulty of obtaining live specimens that will hold back study of the temporal dynamics of such an association. I then plan to explore the spatial distribution of *Fragilariopsis*- and tintinnid-related genes in the *Tara* Oceans metatranscriptomic data, in relationship to the different environmental conditions in which the consortia was observed.

Overall these preliminary results are encouraging, in order to combine the morphological, molecular, functional and ecological characterization of diatom biotic interactions. Successful high throughput sequencing from formol-glutaraldehyde samples enables simultaneous fluorescence microscopy and molecular identification, even though contamination is present. This is promising for further bacterial and eukaryotic molecular identification from ethanol fixed samples. Moreover, even though a large proportion of transcripts remain unassigned both functionally and taxonomically, many steps remain to be tested that are likely to help us gain insight into the metabolic processes involved in diatom biotic interactions. Ultimately, the combination of microfluidics applied to individualized interactions and single cell genomics make it reasonable to believe that the gap between metabolic pathways, cells, and eco-systems is gradually shrinking.

Chapter 6: Conclusions and perspectives

The goal of my thesis was to explore global scale patterns of diatom biotic interactions in the open ocean, in order to understand how they structure planktonic assemblages. By combining both theoretical and experimental procedures, whilst building upon previous empirical knowledge regarding diatom associations, I provide ecosystem level description of the impact of both direct and indirect interactions on assemblages, in relation with a changing environment. The unprecedented extent of the *Tara* Oceans data enables questioning in ways and to a scale that have hardly been at reach for marine microbiologists. Developing new methods, approaches, and interpretation of the data is of utmost importance in order to extract the most relevant and significant meaning of it.

A methodology to crack the case of the "unknown" diatom barcodes is proposed in **Chapter 2**. It is based on a full study of diatom biogeography that has recently been published by Dr. Malviya - former PhD student and mentor in my lab - revealing global diatom distribution and diversity in the sunlit ocean. Very early whilst learning how to deal with the *Tara* Oceans data, I have been intrigued by the proportion of "unknown" sequences in the metabarcoding survey, particularly those that had a low percentage identity with any sequence in NCBI; they represented nearly 50% of the data in terms of richness (unique OTU), and sometimes up to 15% percent of a sample's abundance. Moreover, preliminary co-occurrence networks inferred from *Tara* Oceans data, displayed highly connected nodes barely assigned to "Bacillariophyta", for which further interpretation would reveal itself shaky.

Amongst the top 100 most abundant "unassigned" diatoms, 18 have been assigned just by a careful look at BLAST results in NCBI, revealing that sequences often matched "uncultured stramenopiles" along with a known diatom genera, following which the automated *Tara* Oceans pipeline would keep the last common ancestor as a precaution. In other cases, the V9 portion of the 18S rDNA was simply insufficient to correctly assign the barcodes, and involved the amplification of larger portions of the corresponding rDNA gene, by following a

specifically developed primer design procedure. The assignation of four abundant and ubiquitous diatom barcodes revealed three new genera completely absent from the *Tara* Oceans assignation: *Shionodiscus*, *Thalassiothrix*, *Asteromphalus*. The last barcode was re-assigned to *Pseudo-nitzschia delicatissima* and curiously found in great abundance in the metabarcoding data set.

The Tara Oceans data is so immense that this primer pipeline would deserve a dedicated online service, which could, depending on the barcode, automatically detect the *Tara* sample in which the barcode is the most abundant, and based on all the sequences present in the sample, edit a list of optimal primers to test. The scientists would then order such samples (for which 18S pre-amplification has been done), and proceed to a finer assignation of abundant, poorly annotated sequences. I am currently working with Sarah Romac, close collaborator from the Station Biologique de Roscoff, to design primers for eukaryotic sequences that are "unassigned" and are not even related to a supergroup. Moreover, facing the problem of designing species-specific probes that would simultaneously exclude specific classes of organisms, I was somehow surprised that specialists still rely on manual curation of aligned sequences to design their primers. I have therefore developed another pipeline to target a maximum of sequences from alignment A, whilst excluding a maximum of sequences from alignment B. This was tested *in silico*, but not experimentally and could reveal itself useful in *Tara* Oceans samples, if one wants to design truly universal diatom primers that would exclude all other eukaryotes in environmental samples.

Chapter 2 therefore emphasizes the extent to which diatoms are ubiquitous in the open ocean. My previous work on coral symbiosis and parasitic biological control guided my interest towards diatom interactions for my PhD, in order to understand how biotic associations might impact the planktonic community structure at large spatial scale. Several questions directed my research, as to whether the competitive reputation of diatoms would have a signature at large global scale, if abundant diatoms were key players in the community structure, or if specific co-occurrences were driven by a shared preference for a specific niche. Many answers are provided in **Chapter 3**, demonstrating the fact that diatoms are, along with polycystines, the only supergroup that acts as repulsive segregators and that species co-occurrence is driven by abiotic factors in a minority of cases. It will be

important to confront these results with the extended dataset of 126 stations. Ultimately, I would have wished to obtain a form of decision tree to go towards biological edge annotation. Given an edge between two nodes, we could assign a biotic interaction based on the abundance profile, the size fraction, support from the literature and microscopy, correlation value or betweenness centrality of the node. Despite the recent literature applying graph theory to microbial association networks, many challenges remain such as 1) which procedure to use, 2) obstacles to assess statistical significance of computed interactions, 3) confounding factors, 4) the obvious lack of procedure to analyse constructed networks and, 5) most of all, the difficulty to ecologically interpret the global network structural properties.

Future research directions involve using the *Tara* Oceans samples that have been preserved in fixatives for microscopy in order to quantify biotic interactions and confront this with co-occurrence data. It has also recently been demonstrated that a large bias was inherent to the co-occurrence method used to infer the community structure (Weiss et al., 2016). I have therefore decided to launch a collaboration to apply different co-occurrence methods to the global *Tara* Oceans data set and hence provide more robust statements about diatom pairwise associations. Connecting with the research community working on ecological networks is likely to help us interpret co-occurrence networks. Finally, a research topic of interest to my questions would be the impact of metabolic dependencies on species co-occurrence (Zelezniak et al., 2015) that could be predicted using the *Tara* Oceans metatranscriptomic data.

The *Tara* Oceans data also provides a unique opportunity to integrate genetic, morphological and functional information at ecosystem scale. By focusing on a single type of diatom direct association involving *Fragilariopsis doliolus* and heterotrophic ciliates (**Chapter 4**), I have shown that this consortium occurs worldwide, both in eutrophic and oligotrophic regions, illustrating the success of data-driven decisions, where co-occurrence is a marker of physical association that enables new discovery. However, the interpretation of the true biological consequences of the association remains unknown, as both organisms are not available to culture. Many research directions can help solve this question. Borrowing from physics, modeling fluid dynamics around the consortia can be of use, as it was suggested

that ciliates influence fluxes at the microscale to improve feeding rates. Buoyancy could also be tested, as well as protection from grazing.

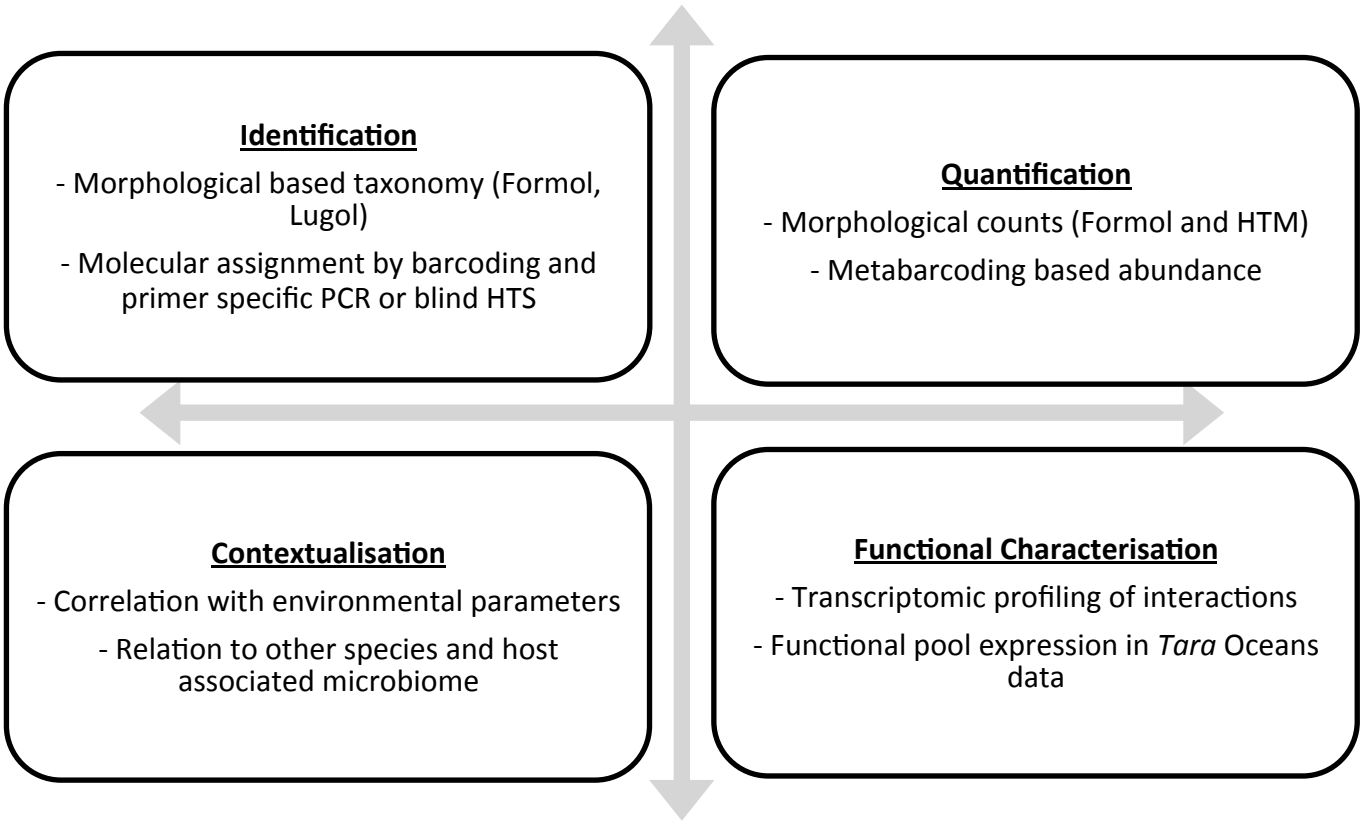


Figure 6.1. Towards a comprehensive characterization of microbial interactions based on the *Tara* Oceans data.

Chapter 5 presents preliminary data that in the future could go a step further to help our understanding of a given consortium (**Figure 6.1**). Bacterial communities are sequenced, in order to understand if the association also serves as a microbial recruitment vector; for example, the tintinnids that feed on bacteria could benefit from those associated the *Fragilariopsis'* biofilm. Transcriptomic data has been extracted from the isolated consortia, with the aim of gaining insight into the metabolic functions carried out by the association, such as exchanges of metabolites, and how they vary through environmental conditions. Our understanding is currently held back by the vast amount of taxonomically and functionally unassigned transcripts, for which better resolved databases and algorithms will help to unlock the situation. Future research directions involve quantification through high throughput microscopy imaging, which could be directly confronted with V9 molecular data. This systemic approach could also be adopted with other novel biotic interactions observed

in the samples, as is done on diatom - parasitic and diatom - epiphytic interactions, using high throughput approaches.

The arrival of high-throughput studies in environmental microbial research has revealed the extent to which our knowledge of marine microbial ecology is scarce. Most environmental studies are based on metabarcoding, the current gold standard for genetic identification, and they regularly report significant proportions of unidentified organisms despite the massive efforts to build up reference databases and control for biases. When a single sample of soil or seawater reveals so many unknown organisms, it makes the empirical knowledge accumulated over the years seem quite limited, most likely because of inclinations towards studying easily cultivable organisms. It nonetheless still gives an incredible estimate of the genetic diversity that is present, even though we often lack any morphological characterization or cultured representatives. In many aspects, the study of interactions in the protistan world seems destined to the same fate. We still lack proper reference databases dedicated to marine interactions, and knowledge is often restricted to robust co-cultures. Co-occurrence networks have flourished in the literature, but have also revealed a lack of interpretability of correlation edges, and even though microbial correlation networks appear as a seductive tool to predict biotic interactions, we still have a long way to go before these methods can reliably and quantitatively propose new associations. However, coupled with empirical knowledge, microbial correlation networks can serve as an excellent way to describe community structure and bring new insights for future research directions of potential key biotic interactions.

Species interactions at the individual level have imprints much larger than the microscopic scales on which they occur. With species-specific mechanistic studies at microscale on the one hand, and ecosystem multispecies and biogeochemical cycle modeling on the other, the two ends of the spectrum have been reasonably explored, though much remains to be discovered between the two approaches. However, we currently lack the analytical and theoretical framework to integrate this into one single picture that can bridge this gap. This will certainly require more interdisciplinarity, increased collaboration and communication

between scientists, to face the challenges of understanding microbial life in a changing ocean.

Works Cited

Agatha S, Laval-Peuto M, Simon P. (2013). The tintinnid lorica, in: Dolan JR, Montagnes DJS, Agatha S, Coats WD, Stoecker DK (Eds.), The biology and ecology of tintinnid ciliates: models for marine plankton. *Wiley-Blackwell, West Sussex, pp. 17-41.*

Aitchison J. (1986). The Statistical Analysis of Compositional Data.

Aitchison J. (1981). A New Approach to Null Correlations of Proportions. *Journal of the International Association for Mathematical Geology Mathematical Geology* 13.2: 175-89.

Albert R, Jeong H, and Barabási A-L. (2000). Error and attack tolerance of complex networks. *Nature* 406.6794: 378-382.

Albright LJ, Yang CZ, and Johnson S. (1993). Sub-lethal Concentrations of the Harmful Diatoms, Chaetoceros Concavicornis and C. Convolutus, Increase Mortality Rates of Penned Pacific Salmon. *Aquaculture* 117.3-4: 215-25.

Alexander H, *et al*. Metatranscriptome analyses indicate resource partitioning between diatoms in the field. *Proceedings of the National Academy of Sciences* 112.17 (2015): E2182-E2190.

Allen AE, *et al*. (2011). Evolution and Metabolic Significance of the Urea Cycle in Photosynthetic Diatoms. *Nature* 473.7346: 203-07.

Amaral-Zettler LA, *et al*. (2009). A Method for Studying Protistan Diversity Using Massively Parallel Sequencing of V9 Hypervariable Regions of Small-Subunit Ribosomal RNA Genes. *PLoS ONE* 4.7.

Amin SA, Parker MS, and Armbrust EV. (2012)**.** Interactions between Diatoms and Bacteria. *Microbiology and Molecular Biology Reviews* 76.3: 667-84.

Araújo MB, Rozenfeld A, Rahbek C, and Marquet PA. (2011). Using Species Co-occurrence Networks to Assess the Impacts of Climate Change. *Ecography* 34.6: 897-908.

Araújo MB, and Rozenfeld A. (2014). The geographic scaling of biotic interactions. *Ecography* 37.5: 406-415.

Arita HT. (2016). Species Co-occurrence Analysis: Pairwise versus Matrix-level Approaches. *Global Ecology and Biogeography*.

Armbrust EV, *et al*. (2004). The Genome of the Diatom Thalassiosira Pseudonana: Ecology, Evolution, and Metabolism. *Science* 306.5693: 79-86.

Armbrust EV. (2009). The life of diatoms in the world's oceans. *Nature* 459.7244: 185-192.

Arumugam M, *et al*. (2011). Enterotypes of the Human Gut Microbiome. *Nature* 474.7353: 666.

Assmy P, Henjes J, Klaas C, and Smetacek V. (2007). Mechanisms Determining Species Dominance in a Phytoplankton Bloom Induced by the Iron Fertilization Experiment EisenEx in the Southern Ocean. *Deep Sea Research Part I: Oceanographic Research Papers* 54.3: 340-62.

Aubert MD, Pesando D, and Gauthier M. (1970). Phenomenes d'antibiose d'origine phytoplanctonique en milieu marin: substances antibactériennes produites par une diatomée Asterionella japonica. *Rev Intern Océanogr Med* 19-20.

Azam F, Fenchel T, Field Jg, Gray Js, Meyer-Reil La, and Thingstad F. (1983). The Ecological Role of Water-Column Microbes in the Sea. *Marine Ecology Progress Series* 10: 257-63.

Azam F. (1998). OCEANOGRAPHY: Microbial Control of Oceanic Carbon Flux: The Plot Thickens. *Science* 280.5364: 694-96.

Bachy C, Dolan JR, López-García P, Deschamps P, and Moreira D. (2012). Accuracy of Protist Diversity Assessments: Morphology Compared with Cloning and Direct Pyrosequencing of 18S RRNA Genes and ITS Regions Using the Conspicuous Tintinnid Ciliates as a Case Study. *The ISME Journal*: 244-55.

Bachy C, Gómez F, López-García P, Dolan JR, and Moreira D. (2012). Molecular Phylogeny of Tintinnid Ciliates (Tintinnida, Ciliophora). *Protist* 163.6: 873-87.

Baldauf SL. (2003). The deep roots of eukaryotes. *Science* 300.5626: 1703-1706.

Balech E. (1962). Tintinnoinea y dinoflagellata del Pacífico: según material de las expediciones Norpac y Downwind del Instituto Scripps de Oceanografía. *Imprenta y Casa Editora Coni*.

Balzano S, *et al.* (2015). Transcriptome analyses to investigate symbiotic relationships between marine protists. *Frontiers in microbiology*.

Ban S, Burns C, *et al.* (1997). The Paradox of Diatom-copepod Interactions. *Marine Ecology Progress Series* 157: 287-93.

Banse K. (2013). Reflections About Chance in My Career, and on the Top-Down Regulated World. *Annual Review of Marine* 5.1: 1-19.

Barberán A, *et al.* (2012). Using network analysis to explore co-occurrence patterns in soil microbial communities. *The ISME journal* 6.2: 343-351.

Barnes RD. (1982). Invertebrate Zoology. Philadelphia, PA: Holt-Saunders International. *ISBN* 0-03-056747-5.

Baselga A, *et al.* (2012). Global patterns in the shape of species geographical ranges reveal range determinants. *Journal of Biogeography* 39.4: 760-771.

Bascompte J, Jordano P, and Olesen JM. (2006). Asymmetric Coevolutionary Networks Facilitate Biodiversity Maintenance. *Science* 312.5772: 431-33.

Bascompte J, and Jordano P. (2007). Plant-Animal Mutualistic Networks: The Architecture of Biodiversity. *Annual Review of Ecology, Evolution, and Systematics* 38.1: 567-93.

Bascompte J. (2009). Disentangling the Web of Life. *Science* 325.5939: 416-19.

Bascompte J. (2010). Structure and Dynamics of Ecological Networks. *Science* 329.5993: 765-66.

Bateman BL, Vanderwal J, Williams SE, and Johnson CN. (2012). Biotic Interactions Influence the Projected Distribution of a Specialist Mammal under Climate Change. *Diversity and Distributions Diversity Distrib.* 18.9: 861-72.

Bergh I, *et al*. (1989). High abundance of viruses found in aquatic environments. *Nature* 340.6233: 467-468.

Behrenfeld MJ, and Falkowski PG. (1997). Photosynthetic rates derived from satellite-based chlorophyll concentration. *Limnology and oceanography* 42.1: 1-20.

Berry D, and Widder S. (2014). Deciphering Microbial Interactions and Detecting Keystone Species with Co-occurrence Networks. *Frontiers in Microbiology*

Biard T, *et al.* (2015). Towards an Integrative Morpho-molecular Classification of the Collodaria (Polycystinea, Radiolaria). *Protist* 166.3: 374-388.

Blaxter M, *et al*. (2005). Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1462: 1935-1943.

Bork P, *et al*. (2015). Tara Oceans studies plankton at planetary scale. *Science* 348.6237: 873-873.

Bollens GR, and Landry M. (2000). Biological Response to Iron Fertilization in the Eastern Equatorial Pacific (IronEx II). II. Mesozooplankton Abundance, Biomass, Depth Distribution and Grazing. *Marine Ecology Progress Series* 201: 43-56.

Borthagaray AI, Arim M, and Marquet PA. (2014). Inferring Species Roles in Metacommunity Structure from Species Co-occurrence Networks. *Proceedings of the Royal Society B: Biological Sciences* 281.1792: 20141425.

Bowler C, *et al*. (2008). The Phaeodactylum Genome Reveals the Evolutionary History of Diatom Genomes. *Nature* 456.7219: 239-44.

Brehm-Stecher BF, and Johnson EA. (2004). Single-cell microbiology: tools, technologies, and applications. *Microbiology and molecular biology reviews* 68.3: 538-559.

Brilli M, *et al.* (2010). The Structural Network Properties of Biological Systems. *Handbook On Biological Networks*. 9-31.

Bronstein JL. (1994). Our current understanding of mutualism. *Quarterly Review of Biology* 31-51.

Brose U, Berlow EL, and Martinez ND. (2005). From Food Webs To Ecological Networks. *Dynamic Food Webs*: 27-36.

Brown M, *et al.* (2009). Microbial community structure in the North Pacific ocean. *The ISME Journal* 3.12: 1374-1386.

Brown TA, Belt ST, Tatarek A, and Mundy CJ. (2014). Source Identification of the Arctic Sea Ice Proxy IP25. *Nature Communications*.

Brusca RC, Richard C, and Gilligan MR. (1983). Tongue replacement in a marine fish (Lutjanus guttatus) by a parasitic isopod (Crustacea: Isopoda). *Copeia* 3: 813-816.

Buchan A, *et al*. (2014). Master recyclers: features and functions of bacteria associated with phytoplankton blooms. *Nature Reviews Microbiology* 12.10: 686-698.

Bucklin A, and Allen LD. MtDNA sequencing from zooplankton after long-term preservation in buffered formalin. *Molecular Phylogenetics and Evolution* 30.3: 879-882.

Calbet A, and Landry MR. (2004). Phytoplankton Growth, Microzooplankton Grazing, and Carbon Cycling in Marine Systems*. Limnology and Oceanography* 49.1: 51-57.

Campbell RG, *et al*. (2009). Mesozooplankton Prey Preference and Grazing Impact in the Western Arctic Ocean. *Deep Sea Research Part II: Topical Studies in Oceanography* 56.17: 1274-289.

Caron DA, *et al*. (2009). Protists are microbes too: a perspective. *The ISME journal* 3.1: 4-12.

Carotenuto Y, *et al*. (2014). Insights into the Transcriptome of the Marine Copepod Calanus Helgolandicus Feeding on the Oxylipin-producing Diatom Skeletonema Marinoi. *Harmful Algae* 31: 153-62.

Carpenter EJ, *et al.* (1999). Extensive bloom of a N_2-fixing diatom/cyanobacterial association in the tropical Atlantic Ocean. . *Marine Ecology Progress Series* 85: 273-283.

Carpenter EJ, and Foster RA. (2002). Marine cyanobacterial symbioses. *Cyanobacteria in symbiosis*. Springer Netherlands 11-17.

Cazelles K, *et al*. (2015). A Theory for Species Co-occurrence in Interaction Networks. *Theoretical Ecology* 9.1: 39-48.

Cermeño P, and Falkowski PG. (2009). Controls on Diatom Biogeography in the Ocean. *Science* 325.5947: 1539-541.

Cermeño P, *et al*. (2015). Continental Erosion and the Cenozoic Rise of Marine Diatoms. *Proceedings of the National Academy of Sciences* 112.14: 4239-244.

Chaffron S, *et al*. (2010). A Global Network of Coexisting Microbes from Environmental and Whole-genome Sequence Data. *Genome Research* 20.7: 947-59.

Chase JM. (2000). Are There Real Differences among Aquatic and Terrestrial Food Webs? *Trends in Ecology & Evolution* 15.10: 408-12.

Chen M, and Liu H. (2011). Experimental Simulation of Trophic Interactions among Omnivorous Copepods, Heterotrophic Dinoflagellates and Diatoms. *Journal of Experimental Marine Biology and Ecology* 403.1-2: 65-74.

Chow CET, *et al*. (2013). Top-down Controls on Bacterial Community Structure: Microbial Network Analysis of Bacteria, T4-like Viruses and Protists. *The ISME Journal* 8.4: 816-29.

Clement M, Posada D, Crandall KA. (2000). TCS: a computer program to estimate gene genealogies. *Mol Ecol* 9: 1657–1660

Cohen AM. (2005). A Survey of Current Work in Biomedical Text Mining. *Briefings in Bioinformatics* 6.1: 57-71.

Coleman A. (1083). Algal Research Progress in Phycological Research. Volume I F. E. Round D. J. Chapman. *BioScience* 33.7: 467.

Connor EF, and Simberloff D. (1979). The assembly of species communities: chance or competition? *Ecology* 60.6: 1132-1140.

Cullen JJ. (1991). Hypotheses to Explain High-nutrient Conditions in the Open Sea. *Limnology and Oceanography* 36.8: 1578-599.

Cusick ME, *et al*. (2009). Literature-curated Protein Interaction Datasets. *Nature Methods* 6.1: 39-46.

D'ippolito G, *et al.* (2002). New Birth-control Aldehydes from the Marine Diatom Skeletonema Costatum: Characterization and Biogenesis. *Tetrahedron Letters* 43.35: 6133-136.

Darriba D, *et al*. (2012). JModelTest 2: More Models, New Heuristics and Parallel Computing. *Nature Methods* 9.8: 772.

Darwin C. *On the Origin of Species, 1859*. Washington Square, NY: New York UP, 1988.

Decelle J, *et al.* (2014). Intracellular Diversity of the V4 and V9 Regions of the 18S rRNA in Marine Protists (Radiolarians) Assessed by High-Throughput Sequencing. *PLoS ONE*

Decelle J, Colin S, and Foster RA. (2015). Photosymbiosis in marine planktonic protists. *Marine Protists*. Springer Japan 465-500.

DeLong EF. (2005). Microbial community genomics in the ocean. *Nature Reviews Microbiology* 3.6: 459-469.

Deng YE, *et al*. (2012). Molecular Ecological Network Analyses. *BMC Bioinformatics* 13.1: 113.

Diamond JM. Assembly of species communities. Á In: Cody, ML and Diamond, JM (eds), Ecology and evolution of communities." (1975).

Dick WA. "Agroecology: Ecological processes in sustainable agriculture."*Journal of Environmental Quality* 28.1 (1999): 354.

Díez B, Pedrós-Alió C, and Massana R. (2001). Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. *Applied and environmental microbiology* 67.7: 2932-2941.

Dolan JR, *et al*. (2002). Microzooplankton Diversity: Relationships of Tintinnid Ciliates with Resources, Competitors and Predators from the Atlantic Coast of Morocco to the Eastern Mediterranean. *Deep Sea Research Part I: Oceanographic Research Papers* 49.7: 1217-232.

Doncheva NT, *et al.* (2012). Topological Analysis and Interactive Visualization of Biological Networks and Protein Structures." *Nature Protocols* 7.4 (2012): 670-85.

Ducklow HW. (1983). Production and Fate of Bacteria in the Oceans. *BioScience* 33.8: 494.

Dunne JA, Williams RJ, and Martinez ND. (2002). Food-web Structure and Network Theory: The Role of Connectance and Size. *Proceedings of the National Academy of Sciences* 99.20: 12917-2922.

Dunne JA, Williams RJ, and Martinez ND. (2002). Network structure and biodiversity loss in food webs: robustness increases with connectance. *Ecology letters* 5.4: 558-567.

Dunne JA, *et al.* (2013). Parasites Affect Food Web Structure Primarily through Increased Diversity and Complexity. *PLoS Biology* 11.6.

D'Ippolito G, *et al*. (2015). Potential of Lipid Metabolism in Marine Diatoms for Biofuel Production." *Biotechnol Biofuels Biotechnology for Biofuels* 8.1: 28.

Edlund MB, and Stoermer EF. (1997). Ecological, Evolutionary, And Systematic Significance Of Diatom Life Histories1. *Journal of Phycology J Phycol* 33.6: 897-918.

Eiler A, Heinrich F, and Bertilsson S. (2011). Coherent Dynamics and Association Networks among Lake Bacterioplankton Taxa. *The ISME Journal* 6.2: 330-42.

Falkowski PG. (1998). Biogeochemical Controls and Feedbacks on Ocean Primary Production. *Science* 281.5374: 200-06.

Falkowski PG. (2002). The Ocean's Invisible Forest. *Scientific American* 287.2: 54-61.

Falkowski PG. (2004). The Evolution of Modern Eukaryotic Phytoplankton. *Science* 305.5682: 354-60.

Falkowski PG, and Knoll AH. *Evolution of Primary Producers in the Sea*. Amsterdam: Elsevier Academic, 2007.

Falkowski PG, Fenchel T, and Delong EF. (2008). The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science* 320.5879: 1034-039.

Falkowski PG, and Raven JA. (2013). Aquatic Photosynthesis.

Faust K, and Raes J. (2012). Microbial interactions: from networks to models. *Nature Reviews Microbiology* 10.8: 538-550.

Faust K, *et al*. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS Comput Biol* 8.7: e1002606.

Feizi S, *et al*. (2013). Network deconvolution as a general method to distinguish direct dependencies in networks. *Nature biotechnology* 31.8: 726-733.

Fernandes LF, and Calixto-Feres M. (2012). Morfologia e distribuição de duas diatomáceas (Bacillariophyta) epizóicas no Brasil. *Acta Botanica Brasilica*, 26(4), 836-841

Field CB. (1998). Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science* 281.5374: 237-40.

Flynn KJ, and Irigoien X. (2009). Aldehyde-induced Insidious Effects Cannot Be Considered as a Diatom Defence Mechanism against Copepods. *Marine Ecology Progress Series* 377: 79-89.

Fogg GE. (1982). Marine plankton. *The biology of cyanobacteria* 4.9: 1-513.

Fol H. (1883). Nouvelle contribution à la connaissance de la famille des Tintinnodea. *Archives des Sciences Physiques et Naturelles* (Genève) 9: 554-578

Fontaine C, *et al*. (2011). The Ecological and Evolutionary Implications of Merging Different Types of Networks. *Ecology Letters* 14.11: 1170-181.

Foster RA and Zehr JP (2006). Characterization of Diatom–cyanobacteria Symbioses on the Basis of NifH, HetR and 16S RRNA Sequences. *Environ Microbiol Environmental Microbiology* 8.11: 1913-925.

Foster RA, *et al*. (2011). Nitrogen fixation and transfer in open ocean diatom–cyanobacterial symbioses. *The ISME journal* 5.9: 1484-1493.

Frank DN, *et al.* (2011). Investigating the Biological and Clinical Significance of Human Dysbioses. *Trends in Microbiology* 19.9: 427-34.

Froneman PW, Pakhomov EA, and Meaton V. (1998). Observations on the association between the diatom, Fragilariopsis doliolus Wallich, and the tintinnid, Salpingella subconica Kafoid. *South African journal of science* 94.5.

Fuhrman J, and Steele JA. (2008). Community Structure of Marine Bacterioplankton: Patterns, Networks, and Relationships to Function*. Aquatic Microbial Ecology* 53: 69-81.

Fuhrman JA. (2009). Microbial Community Structure and Its Functional Implications. *Nature* 459.7244: 193-99.

Fuhrman JA, Cram JA, and Needham DM. (2015). Marine microbial community dynamics and their ecological interpretation. *Nature Reviews Microbiology* 13.3: 133-146.

Furusawa G, *et al*. (2003). Algicidal Activity and Gliding Motility of Saprospira Sp. SS98-5. *Can. J. Microbiol. Canadian Journal of Microbiology* 49.2: 92-100.

Gallina A, *et al.* (2014). The Effect of Polyunsaturated Aldehydes on Skeletonema Marinoi (Bacillariophyceae): The Involvement of Reactive Oxygen Species and Nitric Oxide. *Marine Drugs* 12.7: 4165-187.

Galloway J, *et al.* (2013). A chronology of human understanding of the nitrogen cycle. *Philosophical Transactions of the Royal Society B: Biological Sciences* 368.1621: 20130120.

Gárate-Lizárraga I, and Muñetón-Gómez MS. (2009). Primer registro de la diatomea epibionte Pseudohimantidium pacificum y de otras asociaciones simbióticas en el Golfo de California. *Acta Botanica Mexicana* 88: 31-45.

Gärdes A, *et al.* (2010). Diatom-associated Bacteria Are Required for Aggregation of Thalassiosira Weissflogii. *The ISME Journal* 5.3: 436-45.

Gause HF. (1934). Experimental analysis of Vito Volterra's mathematical theory of the struggle for existence. *Science* 79.2036: 16-17.

Glockner FO, *et al*. (2012). Marine Microbial Diversity and its role in Ecosystem Functioning and Environmental Change. Marine Board Position Paper 17. Calewaert, J.B. and McDonough N. (Eds.).Marine Board-ESF, Ostend, Belgium.

Godhe A, *et al.* (2008). Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. *Applied and environmental microbiology* 74.23 : 7174-7182.

Gómez F, *et al*. (2007). Two high-nutrient low-chlorophyll phytoplankton assemblages: the tropical central Pacific and the offshore Perú-Chile Current." *Biogeosciences* 4.6: 1101-1113.

Gómez F. (2007). On the Consortium of the Tintinnid Eutintinnus and the Diatom Chaetoceros in the Pacific Ocean. *Marine Biology* 151.5: 1899-906.

Gómez F, *et al*. (2011). Solenicola setigera is the first characterized member of the abundant and cosmopolitan uncultured marine stramenopile group MAST-3. *Environmental microbiology* 13.1: 193-202.

Gong, Jun, et al. "Extremely high copy numbers and polymorphisms of the rDNA operon estimated from single cell analysis of oligotrich and peritrich ciliates." *Protist* 164.3 (2013): 369-379.

Gotelli NJ, and Mccabe DJ. (2002). Species Co-Occurrence: A Meta-Analysis of J. M. Diamond's Assembly Rules Model. *Ecology* 83.8: 2091.

Gotelli, Nicholas J., and Werner Ulrich. "The Empirical Bayes Approach as a Tool to Identify Non-random Species Associations." *Oecologia* 162.2 (2009): 463-77.

Gotelli NJ, Graves GR, and Rahbek C. (2010). Macroecological signals of species interactions in the Danish avifauna. *Proceedings of the National Academy of Sciences* 107.11: 5030-5035.

Gouy M, Guindon S, and Gascuel O. (2009). SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Molecular Biology and Evolution* 27.2: 221-24.

Gowing MM, and Garrison DL. (1992). Abundance and Feeding Ecology of Larger Protozooplankton in the Ice Edge Zone of the Weddell and Scotia Seas during the Austral Winter. Deep Sea Research Part A. *Oceanographic Research Papers* 39.5: 893-919.

Gravel D, *et al*. (2011). Trophic theory of island biogeography. *Ecology letters* 14.10: 1010-1016.

Green JL, Bohannan BJM, and Whitaker RJ. (2008). Microbial Biogeography: From Taxonomy to Traits. *Science* 320.5879: 1039-043.

Gsell AS, *et al*. (2013). Temperature Alters Host Genotype-Specific Susceptibility to Chytrid Infection. *PLoS ONE* 8.8.

Gügi B, *et al*. (2015). Diatom-Specific Oligosaccharide and Polysaccharide Structures Help to Unravel Biosynthetic Capabilities in Diatoms. *Marine Drugs* 13.9: 5993-6018.

Guimarães PR, *et al*. (2007). Interaction Intimacy Affects Structure and Coevolutionary Dynamics in Mutualistic Networks. *Current Biology* 17.20: 1797-803.

Guindon S, *et al*. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* 59.3: 307-21.

Haas BJ, *et al*. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*.

Haeckel E. "23. Kunstformen Der Natur / Art Forms of Nature." *Eine Geschichte Der Natur Als Modell Für Formfindung in Ingenieurbau, Architektur Und Kunst - A History of Nature as Model for Design in Engineering, Architecture and Art. Form Follows Nature*.

Hamm CE, *et al*. (2003). Architecture and Material Properties of Diatom Shells Provide Effective Mechanical Protection. *Nature* 421.6925: 841-43.

Hamm C, and Smetacek V. (2007). Armor: why, when, and how. *Evolution of primary producers in the sea* 311-332.

Hansen B, Bjornsen PK, and Hansen PJ. (1994). The Size Ratio between Planktonic Predators and Their Prey. *Limnol. Oceanogr. Limnology and Oceanography* 39.2: 395-403.

Harder T. (2009). Marine Epibiosis: Concepts, Ecological Consequences and Host Defence. Marine and Industrial Biofouling Springer Series on Biofilms: 219-31.

Hardin G. (1960). The competitive exclusion principle. *Science* 131.3409:1292-1297.

Harper HE, and Knoll AH. (1975). Silica, Diatoms, and Cenozoic Radiolarian Evolution. *Geology* 3.4: 175.

Harvey HW, *et al*. (1935). Plankton production and its control. *Journal of the Marine Biological Association of the United Kingdom (New Series)* 20.02: 407-441.

Hasle GR, and Syvertsen EE. (1997). Marine Diatoms. *Identifying Marine Phytoplankton* 5-385

Hawkins BA. (1992). Parasitoid-Host Food Webs and Donor Control. *Oikos* 65.1: 159.

Hebert PDN, Cywinska A, Ball SL, and Dewaard JR. (2003). Biological Identifications through DNA Barcodes. *Proceedings of the Royal Society B: Biological Sciences* 270.1512: 313-21.

Heleno RC, *et al*. (2014). Ecological Networks: Delving into the Architecture of Biodiversity. *Biology Letters* 10.1: 20131000.

Hilton JA, *et al*. (2013). Genomic deletions disrupt nitrogen metabolism pathways of a cyanobacterial diatom symbiont. *Nature communications* 4: 1767.

Hobbie JE, *et al*. (1972). A Study Of The Distribution And Activity Of Microorganisms In Ocean Water1. *Limnology and Oceanography* 17.4: 544-55.

Hobbie JEL, and Daley RJ. (1977). Use of nuclepore filters for counting bacteria by fluorescence microscopy. *Applied and environmental microbiology* 33.5: 1225-1228.

Hoek C, Mann D, and Jahns HM. *Algae: an introduction to phycology*. Cambridge university press, 1995.

Holzmann M, Berney C, Hohenegger J. (2006). Molecular identification of diatom endosymbionts in nummulitid Foraminifera. *Symbiosis* 42, 93–101.

Horner-Devine, *et al*. (2007). A Comparison Of Taxon Co-Occurrence Patterns For Macro- And Microorganisms. *Ecology* 88.6: 1345-353.

Hortal J, *et al*. (2015). Seven Shortfalls That Beset Large-Scale Knowledge of Biodiversity. *Annual Review of Ecology, Evolution, and Systematics* 46.1: 523-49.

Hubbell SP. *The unified neutral theory of biodiversity and biogeography (MPB-32)*. Vol. 32. Princeton University Press, 2001.

Huelsenbeck JP. (2001). Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. *Science* 294.5550: 2310-314.

Hutchinson GE. (1959). Homage to Santa Rosalia or why are there so many kinds of animals? *The American Naturalist* 93.870: 145-159.

Ianora A, *et al*. (2004). Aldehyde Suppression of Copepod Recruitment in Blooms of a Ubiquitous Planktonic Diatom. *Nature* 429.6990: 403-07.

Ings TC, *et al.* (2009). Review: Ecological Networks - beyond Food Webs. *Journal of Animal Ecology* 78.1: 253-69.

Irigoien X, *et al*. (2000). Feeding Selectivity and Egg Production of Calanus Helgolandicus in the English Channel. *Limnology and Oceanography* 45.1: 44-54.

Irigoien X, *et al*. (2002). Copepod hatching success in marine ecosystems with high diatom concentrations. *Nature* 419.6905/ 387-389.

Irigoien X. (2005). Phytoplankton Blooms: A 'loophole' in Microzooplankton Grazing Impact? *Journal of Plankton Research* 27.4. 313-21.

Itoh K. (1970). A consideration on feeding habits of planktonic copepods in relation to the structure of their oral parts. *Bull. Plankton Soc. Japan* 17: 1-10.

Iyer S, *et al*. (2013). Attack robustness and centrality of complex networks. *PloS one* 8.4: e59613.

Janson S, *et al*. (1999). Host specificity in the Richeliaądiatom symbiosis revealed by hetR gene sequence analysis. *Environmental microbiology* 1.5: 431438.

Jarman SN. (2004). Amplicon: software for designing PCR primers on aligned DNA sequences. *Bioinformatics*

Jeong HJ, et al. (2004). Feeding by the Heterotrophic Dinoflagellate Protoperidinium Bipes on the Diatom Skeletonema Costatum. *Aquatic Microbial Ecology* 36: 171-79.

Jónasdóttir S, *et al*. (1998). Role of Diatoms in Copepod Production:good, Harmless or Toxic? *Marine Ecology Progress Series* 172: 305-08.

Jones P, *et al*. (2014). InterProScan 5: genome-scale protein function classification *Bioinformatics.*

Jonsson PR, Johansson M, and Pierce RW. (2004). Attachment to Suspended Particles May Improve Foraging and Reduce Predation Risk for Tintinnid Ciliates. *Limnology and Oceanography* 49.6: 1907-914.

Kaczmarska I, *et al*. (2005). Diversity and Distribution of Epibiotic Bacteria on Pseudo-nitzschia Multiseries (Bacillariophyceae) in Culture, and Comparison with Those on Diatoms in Native Seawater. *Harmful Algae* 4.4: 725-41.

Kanter I, and Kalisky T. (2015). Single cell transcriptomics: methods and applications. *Frontiers in oncology*.

Karsenti E, *et al.* (2011). A holistic approach to marine eco-systems biology. *PLoS biol* 9.10: e1001177.

Katoh K, and Standley DM. (2013). MAFFT: Iterative Refinement and Additional Methods. *Methods in Molecular Biology Multiple Sequence Alignment Methods* 131-46.

Keeling PJ. (2004). Diversity and Evolutionary History of Plastids and Their Hosts. *American Journal of Botany* 91.10: 1481-493.

Keeling PJ, *et al*. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol*12.6: e1001889.

Kennedy K, Hall MW, Lynch MDJ, Moreno-Hagelsieb G, and Neufeld JD. (2014). Evaluating Bias of Illumina-Based Bacterial 16S RRNA Gene Profiles. *Applied and Environmental Microbiology* 80.18: 5717-722.

Kitano H. (2002). Computational systems biology. *Nature* 420.6912: 206-210.

Kiørboe T. A Mechanistic Approach to Plankton Ecology. Princeton: Princeton UP, 2008

Kleppel GS, Holliday DV, and Pieper RE. (1991). Trophic interactions between copepods and microplankton: a question about the role of diatoms. *Limnology and Oceanography* 36.1: 172-178.

Konopka A. (2009). What Is Microbial Community Ecology? *The ISME Journal* 3.11: 1223-230.

Kooistra WH, De Stefano M, Mann DG, and Medlin LK. (2003). The Phylogeny of the Diatoms. *Silicon Biomineralization Progress in Molecular and Subcellular Biology*: 59-97.

Kooistra WH, Gersonde R, Medlin LK, and Mann DG. (2007). The Origin and Evolution of the Diatoms: Their Adaptation to a Planktonic Existence. *Evolution of Primary Producers in the Sea*: 207-49.

Kooistra HWF, Forlani G, and De Stefano M. (2009). Adaptations of Araphid Pennate Diatoms to a Planktonic Existence. *Marine Ecology* 30.1: 1-15.

Kraberg A, Baumann M, and Dürselen CD. *Coastal Phytoplankton: Photo Guide for Northern European Seas*. München: Pfeil, 2010.

Kress W, *et al*. (2015). DNA Barcodes for Ecology, Evolution, and Conservation. *Trends in Ecology & Evolution* 30.1: 25-35.

Kröger N, and Poulsen N. (2008). Diatoms—From Cell Wall Biogenesis to Nanotechnology. *Annu. Rev. Genet. Annual Review of Genetics* 42.1: 83-107.

Kurtz ZD, *et al.* (2015). Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Computational Biology.*

Lafferty KD, and Kuris AM. (2002). Trophic strategies, animal diversity and body size. *Trends in Ecology & Evolution* 17.11: 507-513.

Lalli CM, and Parsons TR. (1993). Biological oceanography: an introduction Butterworth–Heinemann.

Lauritano C, *et al*. (2015). Effects of the Oxylipin-producing Diatom Skeletonema Marinoi on Gene Expression Levels of the Calanoid Copepod Calanus Sinicus. *Marine Genomics* 24: 89-94.

Lazarus D, *et al.* (2014). Cenozoic planktonic marine diatom diversity and correlation to climate change. *PloS one* 9.1: e84857.

Lebour MV. (1922). The food of plankton organisms. *Journal of the Marine Biological Association of the United Kingdom (New Series)* 12.04: 644-677.

Lee SO, *et al*. (2000). Involvement of an extracellular protease in algicidal activity of the marine bacterium Pseudoalteromonas sp. strain A28. *Appl. Environ. Microbiol* 66.10:4334-4339.

Legendre L. (1990). The significance of microalgal blooms for fisheries and for the export of particulate organic carbon in oceans. *Journal of Plankton Research* 12.4:681-699.

Legrand C, Rengefors K, Fistarol GO, and Granéli E. (2003). Allelopathy in Phytoplankton - Biochemical, Ecological and Evolutionary Aspects. *Phycologia* 42.4: 406-19.

Li M, Wang J, and Chen J. (2008). A fast agglomerate algorithm for mining functional modules in protein interaction networks. *International Conference on BioMedical Engineering and Informatics*. Vol. 1. IEEE.

Li C, *et al*. (2016). Predicting Microbial Interactions through Computational Approaches. *Methods* 102: 12-19.

Lidicker WZ. (1979). A Clarification of Interactions in Ecological Systems. *BioScience* 29.8: 475-77.

Lima-Mendez G, Faust K, *et al*. (2015). Determinants of Community Structure in the Global Plankton Interactome. *Science* 348.6237: 1262073

Locey KJ, and Lennon JT. (2016). Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*.

Logares R, *et al*. (2013). Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environmental Microbiology*.

Lomolino MV, *et al*. *Biogeography*. No. QH84 L65 2006. Sunderland, MA: Sinauer Associates, 2006.

López-García P, *et al*. (2001). Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* 409.6820: 603-607.

Lupatini M, *et al*. (2014). Network Topology Reveals High Connectance Levels and Few Key Microbial Genera within Soils. *Frontiers in Environmental Science*.

Ma HW, and Zeng AP. (2003). The Connectivity Structure, Giant Strong Component and Centrality of Metabolic Networks. *Bioinformatics* 19.11: 1423-430.

Mann DG. (1999). The Species Concept in Diatoms. *Phycologia* 38.6: 437-95.

Mann K, and Lazier J. (2013). The Oceans and Global Climate Change: Physical and Biological Aspects. *Dynamics of Marine Ecosystems* 390-422.

MacArthur RH. (1972). Geographical ecology: patterns in the distribution of species. Princeton University Press.

MacArthur RH, and Wilson EO. *Theory of Island Biogeography.(MPB-1)*.Vol. 1. Princeton University Press, 2015.

Macdonald JD. (1869) I.—On the structure of the Diatomaceous frustule, and its genetic cycle. *Journal of Natural History* 3.13: 1-8.

Malviya S, *et al*. (2016). Insights into global diatom distribution and diversity in the world's ocean. *Proceedings of the National Academy of Sciences*

Margalef R. (1979). The Organization of Space. *Oikos* 33.2 (1979): 152.

Marshall SM. "Orr. 4P (1955) The biology of a marine copepod, Calanus finmarchicus (Gunnerus)." *Oliver and Boyd, London*.

Massana R. (2015). Protistan Diversity in Environmental Molecular Surveys. *Marine Protists*.

McCann K. (2007). Protecting biostructure. *Nature* 446.7131: 29-29.

McDonald SM, Sarno D, and Zingone A. (2007). Identifying Pseudo-nitzschia Species in Natural Samples Using Genus-specific PCR Primers and Clone Libraries. *Harmful Algae* 6.6: 849-60

McGill BJ, *et al.* (2006). Rebuilding community ecology from functional traits. *Trends in ecology & evolution* 21.4: 178-185.

McManus GB, and Katz LA. (2009). Molecular and morphological methods for identifying plankton: what makes a successful marriage?. *Journal of Plankton Research* 31.10: 1119-1129.

McManus GB, and Santoferrara LF. (2012) Tintinnids in microzooplankton communities. In Dolan, J.R., Montagnes, DJ.S., Agatha, S., Stoecker, D.K. (eds) The Biology and Ecology of Tintinnid Ciliates: Models for Marine Plankton. Wiley-Blackwell: Oxford, pp 198-213.

Medlin LK, Williams DM, and Sims PA. (1993). The Evolution of the Diatoms (Bacillariophyta). I. Origin of the Group and Assessment of the Monophyly of Its Major Divisions." *European Journal of Phycology* 28.4: 261-75.

Medlin LK and Sims PA. (1993). The transfer of *Pseudoeunotia doliolus* to *Fragilariopsis*. Beihefte zur Nova Hedwigia 106: 323-334.

Medlin LK, Kooistra WHCF, and Schmid A. "MM (2000) A review of the evolution of the diatoms–a total approach using molecules, morphology and geology." *The Origin and Early Evolution of the Diatoms: Fossil, Molecular and Biogeographical Approaches. W. Szafer Institute of Botany, Polish Academy of Sciences, Cracow, Poland*: 13-35.

Menden-Deuer S, and Kiørboe T. (2016). Small Bugs with a Big Impact: Linking Plankton Ecology with Ecosystem Processes. *Journal of Plankton Research J. Plankton Res*.

Menezes AB, *et al.* (2015). Network analysis reveals that bacteria and fungi form modules that correlate independently with soil parameters. *Environmental microbiology* 17.8: 2677-2689.

Meyer PE, Lafitte F, and Bontempi G. (2008). minet: AR/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC bioinformatics* 9.1: 1.

Meyer E, and Weis VM. (2012). Study of cnidarian-algal symbiosis in the "omics" age. *The Biological Bulletin* 223.1: 44-65.

Michels J, Vogt J, and Gorb SN. (2012). Tools for crushing diatoms–opal teeth in copepods feature a rubber-like bearing composed of resilin. *Scientific reports* 2: 465.

Milici M, *et al*. (2016). Co-occurrence analysis of microbial taxa in the Atlantic Ocean reveals high connectivity in the free-living bacterioplankton. *Frontiers in microbiology*.

Miller CB. (2009). A Mechanistic Approach to Plankton Ecology. *Journal of Plankton Research* 31.8: 927-28.

Milns I, Beale C, and Smith VA. (2010). Revealing Ecological Networks Using Bayesian Network Inference Algorithms. *Ecology*.

Miralto A, *et al*. (1999). Embryonic Development in Invertebrates Is Arrested by Inhibitory Compounds in Diatoms. *Marine Biotechnology* 1.4: 401-02.

Mitchell A, *et al*. (2014). The InterPro protein families database: the classification resource after 15 year. *Nucleic Acids Research*.

Modigh M. (2005). Effects of Fixatives on Ciliates as Related to Cell Size. *Journal of Plankton Research* 27.8.

Montoya JM, Pimm SL, and Solé RV. (2006). Ecological Networks and Their Fragility. *Nature* 442.7100: 259-64.

Moon-van der Staay V, Yeo S, De Wachter R, and Vaulot D. (2001). Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* 409.6820: 607-610.

Morales-Castilla I, Matias MG, Gravel D, and Araújo MB. (2015). Inferring Biotic Interactions from Proxies. *Trends in Ecology & Evolution* 30.6: 347-56.

Mordret S, *et al*. (2015). The Symbiotic Life of Symbiodinium in the Open Ocean within a New Species of Calcifying Ciliate (Tiarina Sp.). *The ISME Journal* 10.6: 1424-436.

Morel FMM. (2009). The Biogeochemical Cycles of Trace Metals in the Oceans. *Science* 300.5621: 944-47.

Morel FMM. (2003). The Biogeochemical Cycles of Trace Metals in the Oceans. *Science* 300.5621 :944-47.

Morueta-Holme N, *et al.* (2016). A Network Approach for Inferring Species Associations from Co-occurrence Data. *Ecography*.

Moustafa AB, *et al*. (2009). Genomic Footprints of a Cryptic Plastid Endosymbiosis in Diatoms. *Science* 324.593: 1724-726.

Nagasawa S, and Warren A. (1996). Redescription of Vorticella Oceanica Zacharias, 1906 (Ciliophora: Peritrichia) with Notes on Its Host, the Marine Planktonic Diatom Chaetoceros Coarctatum Lauder,1864. *Hydrobiologia* 337.1-3

Nakayama T, *et al*. (2014). Complete genome of a nonphotosynthetic cyanobacterium in a diatom reveals recent adaptations to an intracellular lifestyle. *Proceedings of the National Academy of Sciences* 111.31: 11407-11412.

Naqvi A, *et al*. (2010). Network-based modeling of the human gut microbiome. *Chemistry & biodiversity* 7.5: 1040.

Nelson DM, *et al.* (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 Global Biogeochemical Cycles* 9.3: 359-72.

Newman MEJ. (2003). The structure and function of complex networks. *SIAM review* 45.2: 167-256.

Newman MEJ. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences* 103.23: 8577-8582.

Not F, *et al*. (2012). Diversity and Ecology of Eukaryotic Marine Phytoplankton. *Advances in Botanical Research Genomic Insights into the Biology of Algae* 1-53.

Ollerton J. "Biological barter": patterns of specialization compared across different mutualisms." *Plant-pollinator interactions: from specialization to generalization. University of Chicago Press, Chicago* (2006): 411-435.

Ootsuka S, Suzaki T, and Horiguchi T. Marine Protists: Diversity and Dynamics. Tokyo: Springer, 2015

Osman RW. (1977). A Community For Communities. Ecology and Evolution of Communities. M. L. Cody and J. M. Diamond, Editors. Belknap Press; Cambridge, Mass. 02138. 1975." *Paleobiology* 3.02: 238-44.

Paracer S, and Ahmadjian V. Symbiosis: an introduction to biological associations Oxford University Press. *New York* (2000).

Parada AE, Needham DM, and Fuhrman JA. (2015). Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environmental microbiology*.

Pascual M and Dunne JA. Ecological Networks: Linking Structure to Dynamics in Food Webs. Oxford University Press, Nov. 2005. ISBN 9780199775057.

Paul C, and Pohnert G. (2011). Interactions of the Algicidal Bacterium Kordia Algicida with Diatoms: Regulated Protease Excretion for Specific Algal Lysis. *PLoS ONE* 6.6.

Pasteur G. (1982). A classification review of mimicry systems. *Annual Review of Ecology and Systematics* 13: 169-199.

Pavillard J. (1913). Observations sur les Diatomées (2e série). *Bulletin de la Société Botanique de France* 60:126-133.

Pavillard J. (1935)**.** Péridiens et diatomées pélagiques recueillis par Alain Gerbault entre les iles Marquises et les ils Galapagos. *Bulletin Institut Océanographique* n°699.

Pavlopoulos GA, *et al*. (2012). Using Graph Theory to Analyze Biological Networks. *BioData Mining* 4.1.

Pellissier L, *et al*. (2013). Combining Food Web and Species Distribution Models for Improved Community Projections. *Ecol Evol Ecology and Evolution* 3.13: 4572-583.

Pernthaler J. (2005). Predation on Prokaryotes in the Water Column and Its Ecological Implications*Nature Reviews Microbiology* 3.7: 537-46.

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

Pesant S, *et al*. (2015). Open Science Resources for the Discovery and Analysis of Tara Oceans Data. *Scientific Data*.

Petrie JR, *et al*. (2010). Metabolic Engineering of Omega-3 Long-chain Polyunsaturated Fatty Acids in Plants Using an Acyl-CoA Δ6-desaturase with ω3-preference from the Marine Microalga Micromonas Pusilla." *Metabolic Engineering* 12.3: 233-40.

Pfitzer E. (1869). Uber den Bau und die Zellteilung der Diatomeen. *Bot. Zeitung* 27: 774-776.

Pianka ER. (1970). On r-and K-selection. *The American Naturalist* 104.940: 592-597.

Piganeau G. *Genomic insights into the biology of algae*. Vol. 64. Academic Press, 2012.

Pimm SL. "Food Webs." (1982).

Platt T, *et al*. (2009). The Phenology of Phytoplankton Blooms: Ecosystem Indicators from Remote Sensing. *Ecological Modelling* 220.21: 3057-069.

Pohnert G. (2000). Wound-Activated Chemical Defense in Unicellular Planktonic Algae. *Angewandte Chemie Angew. Chem. Int. Ed.* 39.23: 4352-354.

Pohnert G. (2005). Diatom/Copepod Interactions in Plankton: The Indirect Chemical Defense of Unicellular Algae. *ChemBioChem* 6.6: 946-59.

Pomeroy LR. (1974). The Ocean's Food Web, A Changing Paradigm. *BioScience* 24.9: 499-504.

Pondaven P, *et al*. (2007). Grazing-induced Changes in Cell Wall Silicification in a Marine Diatom. *Protist* 158.1 :21-28.

Poulson KL, *et al*. (2010). Allelopathic compounds of a red tide dinoflagellate have species-specific and context-dependent impacts on phytoplankton. *Marine Ecology Progress Series* 416: 69-78.

Poulson-Ellestad KL, *et al*. (2014). Metabolomics and Proteomics Reveal Impacts of Chemically Mediated Competition on Marine Plankton. *Proceedings of the National Academy of Sciences* 111.24: 9009-014.

Poulson-Ellestad KL, Mcmillan E, Montoya JP, and Kubanek J. (2014). Are Offshore Phytoplankton Susceptible to Karenia Brevis Allelopathy? *Journal of Plankton Research* 36.5 : 1344-356.

Poulsen N, *et al*. (2014). Isolation and Biochemical Characterization of Underwater Adhesives from Diatoms. *Biofouling* 30.4

Prince EK, *et al*. (2008). Effects of harmful algal blooms on competitors: allelopathic mechanisms of the red tide dinoflagellate Karenia brevis. *Limnology and Oceanography* 53.2: 531-541.

Proulx S, Promislow D, and Phillips P. (2005). Network Thinking in Ecology and Evolution. *Trends in Ecology & Evolution* 20.6

Qin J, *et al.* (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464.7285: 59-65.

Quéré C, *et al*. (2005). Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. *Global Change Biology* 11.11: 2016-2040.

Rabosky DL, and Sorhannus U. (2009). Diversity Dynamics of Marine Planktonic Diatoms across the Cenozoic. *Nature* 457.7226: 183-86.

Raes J, and Bork P. (2008). Molecular eco-systems biology: towards an understanding of community function. *Nature Reviews Microbiology* 6.9: 693-699.

Rai AN, Bergman B, and Rasmussen U, eds. *Cyanobacteria in symbiosis*. Kluwer Academic Pub., 2002.

Rao DV, Pan Y, and Smith SJ. (1995). Allelopathy between Rhizosolenia alata (Brightwell) and the toxigenic Pseudonitzschia pungens f. multiseries (Hasle)." *Harmful Marine Algal Blooms. Lavoisier Intercept Ltd, Paris, France*: 681-686.

Ravasz E, *et al*. (2002). Hierarchical organization of modularity in metabolic networks. *Science* 297.5586: 1551-1555.

Rezende EL, *et al*. (2007). Non-random Coextinctions in Phylogenetically Structured Mutualistic Networks. *Nature* 448.7156: 925-28.

Ribalet F, *et al.* (2014). Phytoplankton cell lysis associated with polyunsaturated aldehyde release in the Northern Adriatic Sea. *PloS one* 9.1: e85947.

Rose JM, and Caron DA. (2007). Does Low Temperature Constrain the Growth Rates of Heterotrophic Protists? Evidence and Implications for Algal Blooms in Cold Waters. *Limnol. Oceangr. Limnology and Oceanography* 52.2: 886-95.

Round FE, *et al*. *Diatoms: biology and morphology of the genera*. Cambridge University Press, 1990.

Round FE. (1991). Diatoms in River Water-monitoring Studies. *J Appl Phycol Journal of Applied Phycology* 3.2: 129-45.

Ruggiero MV, *et al*. (2015). Diversity and Temporal Pattern of Pseudo-nitzschia Species (Bacillariophyceae) through the Molecular Lens. *Harmful Algae* 42

Runge JA. (1988). Should we expect a relationship between primary production and fisheries? The role of copepod dynamics as a filter of trophic variability. *Hydrobiologia* 167.1: 61-71.

Rusch DB, *et al*. (2007). The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biology*.

Ianora A, Miralto A, and Poulet SA. (1999). Are Diatoms Good or Toxic for Copepods? *Marine Ecology Progress Series.* 177: 305-08.

Lynn SM. (1967). On the origin of mitosing cells. *Journal of theoretical biology*14.3: 225-IN6.

Santos AMD, *et al*. (2015). TcsBU: A Tool to Extend TCS Network Layout and Visualization. *Bioinformatics* 32.4

Sazhin AF, *et al*. (2007). The Colonization of Two Phaeocystis Species (Prymnesiophyceae) by Pennate Diatoms and Other Protists: A Significant Contribution to Colony Biomass. *Phaeocystis, Major Link in the Biogeochemical Cycling of Climate-relevant Elements.*

Schindelin J, *et al.* (2012). Fiji: An Open-source Platform for Biological-image Analysis. *Nature Methods* 9.7: 676-82.

Schnepf E, Elbrächter M. 1999 Dinophyte chloroplasts and phylogeny—a review. Grana 38, 81–97

Scholz B, *et al*. (2016). Zoosporic Parasites Infecting Marine Diatoms – A Black Box That Needs to Be Opened. *Fungal Ecology* 19: 59-76.

Schulz MH, *et al.* (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*.

Schwab C, *et al.* (2014). Longitudinal Study of Murine Microbiota Activity and Interactions with the Host during Acute Inflammation and Recovery. *The ISME Journal* 8.5: 1101-114.

Shannon P, *et al.* (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* 13.11: 2498-2504.

Sharp JH, Underhill PA, and Hughes DJ. (1979). Interaction (Allelopathy) Between Marine Diatoms: Thalassiosira Pseudonana And Phaeodactylum Tricornutum. *Journal of Phycology J Phycol* 15.4: 353-62.

Shatwell T, Köhler J, and Nicklisch A. (2013). Temperature and photoperiod interactions with silicon-limited growth and competition of two diatoms. *Journal of plankton research*: fbt058.

Sherr E, and Sherr B. (1988). Role of Microbes in Pelagic Food Webs: A Revised Concept. *Limnol. Oceangr. Limnology and Oceanography* 33.5: 1225-227.

Sherr E, and Sherr B. (2007). Heterotrophic Dinoflagellates: A Significant Component of Microzooplankton Biomass and Major Grazers of Diatoms in the Sea. *Marine Ecology Progress Series Mar. Ecol. Prog. Ser.* 352: 187-97.

Sherr E, and Sherr B. (2009). Capacity of Herbivorous Protists to Control Initiation and Development of Mass Phytoplankton Blooms. *Aquatic Microbial Ecology* 57: 253-62.

Shih PM, and Matzke NJ. (2013). Primary endosymbiosis events date to the later Proterozoic with cross-calibrated phylogenetic dating of duplicated ATPase proteins. *Proceedings of the National Academy of Sciences* 110.30: 12355-12360.

Sieburth JM, Smetacek V, and Lenz J. (1978). Pelagic Ecosystem Structure: Heterotrophic Compartments of the Plankton and Their Relationship to Plankton Size Fractions. *Limnol. Oceangr. Limnology and Oceanography* 23.6: 1256-263.

Silverman MS, Davis I, and Pillai DR. (2010). Success of Self-Administered Home Fecal Transplantation for Chronic Clostridium Difficile Infection. *Clinical Gastroenterology and Hepatology* 8.5: 471-73.

Sims PA, *et al*. (2006). Evolution of the Diatoms: Insights from Fossil, Biological and Molecular Data. *Phycologia* 45.4: 361-402.

Sison-Mangus MP, *et al*. (2014). Host-specific adaptation governs the interaction of the marine diatom, Pseudo-nitzschia and their microbiota. *The ISME journal* 8.1: 63-76.

Smetacek V. (1998). Biological oceanography: diatoms and the silicate factor. *Nature* 391.6664: 224-225.

Smetacek V. (1999). Diatoms and the Ocean Carbon Cycle. *Protist* 150.1: 25-32.

Smetacek V. (2001). A Watery Arms Race. *Nature* 411.6839: 745.

Smetacek V. (2012). Making Sense of Ocean Biota: How Evolution and Biodiversity of Land Organisms Differ from That of the Plankton. *Journal of Biosciences* 37.4: 589-607.

Sogin ML, *et al.* (2006). Microbial diversity in the deep sea and the underexplored « rare biosphere ». *Proceedings of the National Academy of Sciences*.

Sommer U, *et al.* (2012). Beyond the Plankton Ecology Group (PEG) Model: Mechanisms Driving Plankton Succession. *Annu. Rev. Ecol. Evol. Syst. Annual Review of Ecology, Evolution, and Systematics* 43.1: 429-48.

Stahl DA, *et al.* (1984). Analysis of hydrothermal vent-associated symbionts by ribosomal RNA sequences. *Science* 224.4647: 409-411.

Stanish LF, *et al.* (2012). Bacteria and Diatom Co-occurrence Patterns in Microbial Mats from Polar Desert Streams. *Environmental Microbiology* 15.4: 1115-131.

Stecher B, Berry D, and Loy A. (2013). Colonization Resistance and Microbial Ecophysiology: Using Gnotobiotic Mouse Models and Single-cell Technology to Explore the Intestinal Jungle. *FEMS Microbiology Reviews FEMS Microbiol Rev* 37.5: 793-829.

Steele JA, *et al.* (2011). Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *The ISME journal* 5.9: 1414-1425.

Stone L, and Roberts A. (1992). Competitive Exclusion, or Species Aggregation? *Oecologia* 91.3: 419-24.

Strom SL, Macri EL, and Olson B. (2007). Microzooplankton Grazing in the Coastal Gulf of Alaska: Variations in Top-down Control of Phytoplankton. *Limnol. Oceanogr. Limnology and Oceanography* 52.4: 1480-494.

Strom SL. (2008). Microbial Ecology of Ocean Biogeochemistry: A Community Perspective. *Science* 320.5879: 1043-045.

Suttle CA. (2007). Marine viruses—major players in the global ecosystem. *Nature Reviews Microbiology* 5.10: 801-812.

Taberlet P, *et al*. (2012). Towards Next-generation Biodiversity Assessment Using DNA Metabarcoding. *Molecular Ecology* 21.8: 2045-050.

Talavera G, and Castresana J. (2007). Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. *Systematic Biology* USYB 56.4

Takana Y, Hansen G, Fujita D, Horiguchi T. (2008). Serial replacement of diatom endosymbionts in two freshwater dinoflagellates, Peridiniopsis spp. (Peridiniales, Dinophyceae). *Phycologia* 47,

Tammilehto A, *et al*. (2015). Induction of domoic acid production in the toxic diatom Pseudo-nitzschia seriata by calanoid copepods. *Aquatic Toxicology* 159:52-61.

Teeling H, *et al*. (2012). Substrate-Controlled Succession of Marine Bacterioplankton Populations Induced by a Phytoplankton Bloom. *Science* 336.6081: 608-11.

Thebault E, and Fontaine C. (2010). Stability of Ecological Communities and the Architecture of Mutualistic and Trophic Networks. *Science* 329.5993: 853-56.

Theriot EC. (2010). A Preliminary Multigene Phylogeny of the Diatoms (Bacillariophyta): Challenges for Future Research. *Plant Ecology and Evolution* 143.3: 278-96

Untergasser A, *et al*. (2012). Primer3--new capabilities and interfaces *Nucleic Acids Research*

Valentini A, Pompanon F, and Taberlet P. (2009). DNA Barcoding for Ecologists. *Trends in Ecology & Evolution* 24.2: 110-17.

Vargas CD, *et al.* (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science.*

Vázquez DP, and Aizen MA. (2004). Asymmetric specialization: a pervasive feature of plant–pollinator interactions. *Ecology* 85.5: 1251-1257.

Veen FJF, *et al*. (2008). Food Web Structure of Three Guilds of Natural Enemies: Predators, Parasitoids and Pathogens of Aphids. *Journal of Animal Ecology* 77.1: 191-200.

Verity PG, and Villareal TA. (1986). The Relative Food Value of Diatoms, Dinoflagellates, Flagellates, and Cyanobacteria for Tintinnid Ciliates. Archiv Für Protistenkunde 131.1-2.

Villareal TA. (1994). Widespread occurrence of the Hemiaulus-cyanobacterial symbiosis in the southwest North Atlantic Ocean. *Bulletin of Marine Science* 54.1:1-7.

Vrieling EG, *et al*. (1999). Diatom Silicon Biomineralization as an Inspirational Source of New Approaches to Silica Production. *Journal of Biotechnology* 70.1-3: 39-51.

Wallace AR. Tropical nature, and other essays. Macmillan and Company,1878.

Wallich GC. (1860). On the siliceous organisms found in the digestive cavities of the Salpae, and their relation to the flint nodules of the Chalk Formation. *Transactions of the Microscopical Society of London, New Series*

Wang Z, Gerstein M, and Snyder M, (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*

Wasik A, Mikolajczyk E, and Ligowski R. (1996). Agglutinated Loricae of Some Baltic and Antarctic Tintinnina Species (Ciliophora). *Journal of Plankton Research* 18.10: 1931-940.

Weiss S, *et al*. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME journal*.

West J, Round FE, Crawford RM, and Mann DG. (1991). The Diatoms: Biology & Morphology of the Genera. *Taxon* 40.1: 156.

Wetherbee R, *et al*. (1998). Minireview-The First Kiss: Establishment And Control Of Initial Adhesion By Raphid Diatoms. *Journal of Phycology* 34.1.

Whitman WB, *et al*. (1998). Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences*95.12.6578-6583.

Whittaker RH. (1972). Evolution and measurement of species diversity. *Taxon*: 213-251.

Wilhelm SW, and Matteson AR. (2008). Freshwater and marine virioplankton: a brief overview of commonalities and differences. *Freshwater Biology* 53.6: 1076-1089.

Williams RJ, Howe A, and Hofmockel KS. (2014). Demonstrating Microbial Co-occurrence Pattern Analyses within and between Ecosystems. *Frontiers in Microbiology*.

Willis RJ. *The History of Allelopathy*. Dordrecht: Springer, 2007.

Wootton JT, and Emmerson M. (2005). Measurement of interaction strength in nature." *Annual Review of Ecology, Evolution, and Systematics*: 419-444.

Yang W, Lopez PJ, and Rosengarten G. (2011). Diatoms: Self Assembled Silicananostructures, and Templates for Bio/chemical Sensors and Biomimetic Membranes." *The Analyst Analyst* 136.1: 42-53.

Yeung LY, *et al*. (2012). Impact of diatom-diazotroph associations on carbon export in the Amazon River plume. *Geophysical Research Letters* 39.18.

Zagoskin MV, *et al.* (2014). Phylogenetic information content of Copepoda ribosomal DNA repeat units: ITS1 and ITS2 impact. *BioMed research international*.

Zelezniak A, *et al*. (2015). Metabolic Dependencies Drive Species Co-occurrence in Diverse Microbial Communities. *Proceedings of the National Academy of Sciences* 112.20: 6449-454.

Zerbino DR, and Birney E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*.

Zhang T, *et al*. (2015). Diversity and distribution of fungal communities in the marine sediments of Kongsfjorden, Svalbard (High Arctic). *Scientific Reports*

Zimmermann J, *et al*. (2014). Taxonomic reference libraries for environmental barcoding: a best practice example from diatom research. *PloS one* 9.9: e108793.

Zinger L, Lucie, Gobet A, and Pommier T. (2012). Two decades of describing the unseen majority of aquatic microbial diversity. *Molecular Ecology* 21.8:1878-1896.

Annexes

A. The contribution of genomics to chart microbial diversity

Massively parallel DNA sequencing made it possible to explore microbial genetic diversity of environmental samples both qualitatively and quantitatively.

Barcoding. Morphological studies have underpinned organisms classification for many decades, based on successive diagnostic features. In recent years, molecular markers (barcodes) have been used for phylogenetic analyses to elucidate the evolutionary history of living organisms, based on the - oversimplified - idea that the further apart two organisms are evolutionarily, the bigger the differences between their genomic sequences. In the strict sense, a DNA barcode is a short sequence taken from a standardized portion of the genome that can be used to identify species, very similar to the barcodes used in supermarkets to equate products with prices. The specific sequence is chosen based on a set of important criteria summarized in Valentini et al., 2009 that includes high **intra-species** similarity balanced with sufficient **inter-species** dissimilarity, standardized amongst different taxonomic groups, containing enough phylogenetic information.

One such widely used example of a DNA region for reconstructing phylogenies are the genes encoding the ribosomal RNA subunits, rDNAs. rDNAs encode the RNA components of the ribosome (rRNAs) and form two subunits, the large subunit (LSU) and the small subunit (SSU). In most eukaryotes, the 18S rRNA is the small ribosomal subunit, and the large subunit contains three rRNA species (5S, 5.8S, 28S in mammals and 25S rRNA in plants). The rRNA encoding genes are typically organized in clusters and are separated by internal transcribed spacers (ITS1 and ITS2), and an intergenic spacer (**Figure A1**).

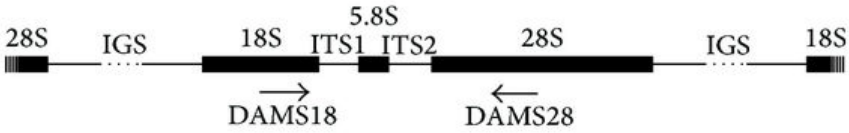


Figure A1. Eukaryotic structure of the 18sRNA.
(a) Typical organisation of tandemly repeated rDNA clusters in eukaryotes. 18S, 5.8S, and 28S ribosomal RNA-encoding genes; ITS1 and ITS2 internal transcribed regions; IGS intergenic regions (Zagoskin, 2014).

Metabarcoding. The impact of barcodes as a molecular tool goes beyond a better-resolved phylogeny of known species, and serves for conservation, species discovery, ecological forensic and community ecology (Kress et al., 2015). Thanks to the arrival of next generation sequencing, the 18S rDNA has become a versatile tool for eukaryotic community ecology through what is now known as *metabarcoding*, "meta" referring to the fact that barcoding is done on environmental samples. More formally, "DNA metabarcoding refers to the automated identification of multiple species from a single bulk sample containing entire organisms or from a single environmental sample containing degraded DNA (soil, water, faeces, etc.)" (Taberlet et al., 2012). A suitable barcode for metabarcoding studies should, amongst other things (i) correspond to a gene region nearly identical among individuals of the same species, but different between species (ii) be suitable for all chosen taxonomic groups, (iii) allow taxonomic assignment, and (iv) should permit the definition of taxonomic levels (Valentini et al., 2009). DNA sequences from complex mixtures of organisms representing different species are obtained through DNA sequencing that localize and simultaneously recover the taxonomic marker gene from the majority of individuals in the sample. Based on sequence similarity, amplicons are then clustered to form Operational Taxonomic Units - OTUs - (Blaxter et al., 2005), a pragmatic proxy to define microbial species. The taxonomic assignment of DNA barcodes from a given environmental sample is done by comparing the amplified sequence with reference databases, typically obtained from organisms grown in culture (**Figure A2**).

A periodic concern with metabarcoding studies is the incomplete and non-representative reference databases used to assign amplified barcodes. It was estimated that Earth is home to as many as one trillion microbial species ($10^{\wedge 12}$) reducing our knowledge of microbial

diversity to 0.001% (Locey et al., 2016) and it is unrealistic to think that cultivating this diversity is the solution to the problem. For relatively well-studied groups such as diatoms, it was shown that the short V9 portion of the 18S rDNA sequence, a 130 base pair long sequence used in metabarcoding studies, does not always contain sufficient phylogenetic information to discriminate between species (Zimmerman et al., 2014). Obtaining larger DNA fragments including other molecular markers could help solve this issue.

An additional technology-related problem concerns the primer specificity, particularly well illustrated by the “Fuhrman primers”, which revealed that a whole class of prokaryotes remained under sequenced based on the falsely universal primers used for the Earth Microbiome Project (Parada et al., 2015). Amplification and sequencing error, as well as clustering (or absence of clustering) of amplicons could furthermore bias the diversity estimates, often overestimating the total amount of individual sequences present in the sample. Beyond diversity estimates, it appears that evaluating abundance of organisms based on high throughput sequencing is prone to error. Indeed, a single diatom cell can display multiple copies of the 18S rDNA, ranging from 61 to 36,896 copies (Godhe et al., 2008). Recent comparative studies which confront molecular diversity and abundance estimates with morphological data in diatoms (Zimmerman et al., 2014), ciliates (Bachy et al., 2012; Gong et al., 2013) and collodarians (Biard et al., 2015) help solve these concerns.

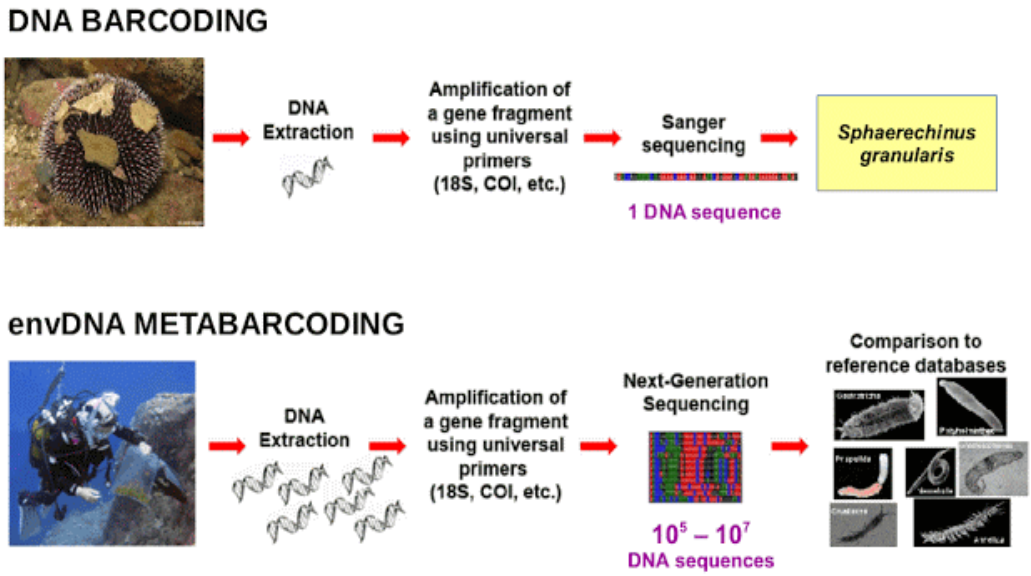


Figure A2. Barcoding versus metabarcoding.
Biological identifications through DNA barcodes (Hebert et al., 2003).

B. Defining biotic interactions

There are many ways to describe biotic interactions- by their type (antagonistic or mutualistic), their strength (weak or strong), their specialization (specialists or generalists) - though it is, in practical, difficult to make microbial interactions fit in one box because many of them are still not mechanistically understood. The words interaction and association are used interchangeably in the dissertation and can be primarily classified in two distinct groups : mutualism and antagonism.

Mutualism involves the exchange of goods and services amongst two species, which become mutualistic partners. Each partner receives a benefit from the interaction, but this generally has a cost. The benefit is not always equal and in any case, species do not behave altruistically. Instead, the benefit is considered as an unintended consequence of the interaction, by which species pursue their own selfish interest (Bronstein, 1994). Emblematic examples in the terrestrial realm are represented by flowering plants and animal pollinators (Ollerton, 2006), or acacia trees and the ants that live in them and protect them in return, or between plants and fungal species that form mycorrhizae.

Antagonism, on the other hand, is an association in which one organism gains benefit at the expense of the other. In predation, one bigger organism often captures biomass from a smaller one and kills it. In parasitism, the smaller parasite will benefit food and shelter from a bigger host but will not kill it, contrary to parasitoids that kill the host. For instance, the *Lithognathus* fish is parasitized by the *Cymothoa exigua* crustacean, that replaces the fish's tongue to feed on its blood and mucus, without apparent damage to the host (Brusca et al., 1983). Hosts and parasites coevolve, shaping the evolutionary arms race, in which short generation time of the parasite generally confers an advantage with quicker adaptation (Dunne et al., 2013).

A fine-tuned classification of biotic interactions can be based on the combination of the relative positive, negative, or neutral effects that the organisms have on each other **(Figure B1)**.

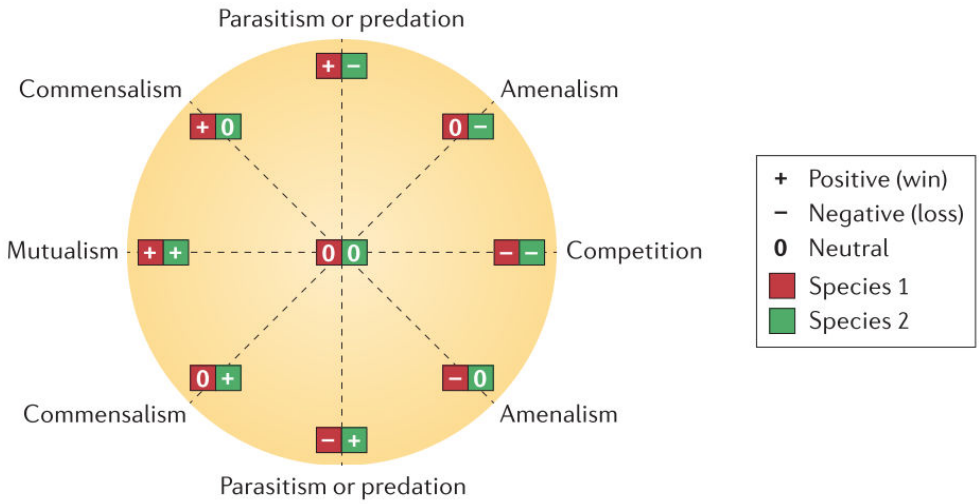


Figure B1. Summary of ecological interactions between different species.
The wheel display by Lidicker has been adapted to summarize all possible pairwise interactions. For each interaction partner, there are three possible outcomes: positive (+), negative (-) and neutral (0). For instance, in parasitism, the parasite benefits from the relationship (+) whereas the host is harmed (-); this relationship is thus represented by the symbol pair +-(Faust and Raes, 2012).

Each of these forms can be further (non-exhaustively) characterized by:

- The degree of dependence: is the interaction obligate or facultative? If obligate, species totally rely on one another for goods and services, such as obligate parasites that depend on their host to complete their life cycle. If facultative, one of the partners can be replaced by another species without affecting the benefit for the other partner(s) (Wootton et al., 2005).
- The degree of specificity: is the interaction between pairs of species (specialists), or pairs of groups (generalists)? Specific mutualism between two species is rare (fig plants, and fig wasps), whereas non-generalist interactions are more common, e.g., whereby honey bees are known to visit the flowers of multiple plant species. Such phenomena lead to highly interconnected networks of plant-pollinator interactions (Vazquez, 2004).

- The degree of physical associations: are the partners physically close when they interact? They are exhabitational, when species such as pollinators live separately from the plants they interact with or ectoparasites that live on the skin of their host. But they are defined as being inhabitational if the partners live with one another (Dick, 1999).

Interspecies interactions can be hard to observe in situ, especially in communities of micro-organisms, and most of our understanding today comes from terrestrial environments, primarily through studies of plant-parasites, plant-pollinator or macroorganism predation (Bascompte, 2009).

C. Co-authored manuscript 1: de Vargas et al, 2015

De Vargas et al., 2015:

I was in charge of the initial version of **Figure 5.C** for the abundance distribution of parasites leading to the observation that putative parasite metabarcodes decreased in the microplanktonic size fractions, but increased again in the mesoplankton. I had initially computed the richness and abundance distribution for each taxonomic group across all combinations of fraction and depth, and significance in distribution difference of each barcode in various fractions was tested with Kruskal Wallis and Wilcox pairwise comparison.

OCEAN PLANKTON

Eukaryotic plankton diversity in the sunlit ocean

Colomban de Vargas,[1,2]*† Stéphane Audic,[1,2]† Nicolas Henry,[1,2]† Johan Decelle,[1,2]† Frédéric Mahé,[3,1,2]† Ramiro Logares,[4] Enrique Lara,[5] Cédric Berney,[1,2] Noan Le Bescot,[1,2] Ian Probert,[6,7] Margaux Carmichael,[1,2,8] Julie Poulain,[9] Sarah Romac,[1,2] Sébastien Colin,[1,2,8] Jean-Marc Aury,[9] Lucie Bittner,[10,11,8,1,2] Samuel Chaffron,[12,13,14] Micah Dunthorn,[3] Stefan Engelen,[9] Olga Flegontova,[15,16] Lionel Guidi,[17,18] Aleš Horák,[15,16] Olivier Jaillon,[9,19,20] Gipsi Lima-Mendez,[12,13,14] Julius Lukeš,[15,16,21] Shruti Malviya,[8] Raphael Morard,[22,1,2] Matthieu Mulot,[5] Eleonora Scalco,[23] Raffaele Siano,[24] Flora Vincent,[13,8] Adriana Zingone,[23] Céline Dimier,[1,2,8] Marc Picheral,[17,18] Sarah Searson,[17,18] Stefanie Kandels-Lewis,[25,26] *Tara* Oceans Coordinators‡ Silvia G. Acinas,[4] Peer Bork,[25,27] Chris Bowler,[8] Gabriel Gorsky,[17,18] Nigel Grimsley,[28,29] Pascal Hingamp,[30] Daniele Iudicone,[23] Fabrice Not,[1,2] Hiroyuki Ogata,[31] Stephane Pesant,[32,22] Jeroen Raes,[12,13,14] Michael E. Sieracki,[33,34] Sabrina Speich,[35,36] Lars Stemmann,[17,18] Shinichi Sunagawa,[25] Jean Weissenbach,[9,19,20] Patrick Wincker,[9,19,20]* Eric Karsenti[26,8]*

Marine plankton support global biological and geochemical processes. Surveys of their biodiversity have hitherto been geographically restricted and have not accounted for the full range of plankton size. We assessed eukaryotic diversity from 334 size-fractionated photic-zone plankton communities collected across tropical and temperate oceans during the circumglobal *Tara* Oceans expedition. We analyzed 18*S* ribosomal DNA sequences across the intermediate plankton-size spectrum from the smallest unicellular eukaryotes (protists, >0.8 micrometers) to small animals of a few millimeters. Eukaryotic ribosomal diversity saturated at ~150,000 operational taxonomic units, about one-third of which could not be assigned to known eukaryotic groups. Diversity emerged at all taxonomic levels, both within the groups comprising the ~11,200 cataloged morphospecies of eukaryotic plankton and among twice as many other deep-branching lineages of unappreciated importance in plankton ecology studies. Most eukaryotic plankton biodiversity belonged to heterotrophic protistan groups, particularly those known to be parasites or symbiotic hosts.

The sunlit surface layer of the world's oceans functions as a giant biogeochemical membrane between the atmosphere and the ocean interior (*1*). This biome includes plankton communities that fix CO_2 and other elements into biological matter, which then enters the food web. This biological matter can be remineralized or exported to the deeper ocean, where it may be sequestered over ecological to geological time scales. Studies of this biome have typically focused on either conspicuous phyto- or zooplankton at the larger end of the organismal size spectrum or microbes (prokaryotes and viruses) at the smaller end. In this work, we studied the taxonomic and ecological diversity of the intermediate size spectrum (from 0.8 μm to a few millimeters), which includes all unicellular eukaryotes (protists) and ranges from the smallest protistan cells to small animals (*2*). The ecological biodiversity of marine planktonic protists has been analyzed using Sanger (*3*–*5*) and high-throughput (*6*, *7*) sequencing of mainly ribosomal DNA (rDNA) gene markers, on relatively small taxonomic and/or geographical scales, unveiling key new groups of phagotrophs (*8*), parasites (*9*), and phototrophs (*10*). We sequenced 18*S* rDNA metabarcodes up to local and global saturations from size-fractionated plankton communities sampled systematically across the world tropical and temperate sunlit oceans.

A global metabarcoding approach

To explore patterns of photic-zone eukaryotic plankton biodiversity, we generated ~766 million raw rDNA sequence reads from 334 plankton samples collected during the circumglobal *Tara* Oceans expedition (*11*). At each of 47 stations, plankton communities were sampled at two water-column depths corresponding to the main hydrographic structures of the photic zone: subsurface mixed-layer waters and the deep chlorophyll maximum (DCM) at the top of the thermocline. A low-shear, nonintrusive peristaltic pump and plankton nets of various mesh sizes were used on board *Tara* to sample and concentrate appropriate volumes of seawater to theoretically recover complete local eukaryotic biodiversity from four major organismal size fractions: piconanoplankton (0.8 to 5 μm), nanoplankton (5 to 20 μm), microplankton (20 to 180 μm), and mesoplankton (180 to 2000 μm) [see (*12*) for detailed *Tara* Oceans field sampling strategy and protocols].

We extracted total DNA from all samples, polymerase chain reaction (PCR)–amplified the hypervariable V9 region of the nuclear gene that encodes 18*S* rRNA (*13*), and generated an average of 1.73 ± 0.65 million sequence reads (paired-end Illumina) per sample (*11*). Strict bioinformatic quality control led to a final data set of 580 million reads, of which ~2.3 million were distinct,

[1]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [2]Sorbonne Universités, Université Pierre et Marie Curie (UPMC) Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [3]Department of Ecology, University of Kaiserslautern, Erwin-Schroedinger Street, 67663 Kaiserslautern, Germany. [4]Department of Marine Biology and Oceanography, Institute of Marine Science (ICM)–Consejo Superior de Investigaciones Científicas (CSIC), Passeig Marítim de la Barceloneta 37-49, Barcelona E08003, Spain. [5]Laboratory of Soil Biology, University of Neuchâtel, Rue Emile-Argand 11, 2000 Neuchâtel, Switzerland. [6]CNRS, FR2424, Roscoff Culture Collection, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [7]Sorbonne Universités, UPMC Paris 06, FR 2424, Roscoff Culture Collection, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [8]Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), and Inserm U1024, and CNRS UMR 8197, Paris, F-75005 France. [9]Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91000 Evry, France. [10]CNRS FR3631, Institut de Biologie Paris-Seine, F-75005, Paris, France. [11]Sorbonne Universités, UPMC Paris 06, Institut de Biologie Paris-Seine, F-75005, Paris, France. [12]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [13]Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. [14]Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [15]Institute of Parasitology, Biology Centre, Czech Academy of Sciences, Branišovská 31, 37005 České Budějovice, Czech Republic. [16]Faculty of Science, University of South Bohemia, Branišovská 31, 37005 České Budějovice, Czech Republic. [17]CNRS, UMR 7093, Laboratoire d'Océanographie de Villefranche-sur-Mer (LOV), Observatoire Océanologique, F-06230, Villefranche-sur-Mer, France. [18]Sorbonne Universités, UPMC Paris 06, UMR 7093, LOV, Observatoire Océanologique, F-06230, Villefranche-sur-Mer, France. [19]CNRS, UMR 8030, CP5706, Evry, France. [20]Université d'Evry, UMR 8030, CP5706, Evry, France. [21]Canadian Institute for Advanced Research, 180 Dundas Street West, Suite 1400, Toronto, Ontario M5G 1Z8, Canada. [22]MARUM, Center for Marine Environmental Sciences, University of Bremen, 28359 Bremen, Germany. [23]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [24]Ifremer, Centre de Brest, DYNECO/Pelagos CS 10070, 29280 Plouzané, France. [25]Structural and Computational Biology, European Molecular Biology Laboratory (EMBL), Meyerhofstraße 1, 69117 Heidelberg, Germany. [26]Directors' Research, EMBL, Meyerhofstraße 1, 69117 Heidelberg, Germany. [27]Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany. [28]CNRS UMR 7232, Biologie Intégrative des Organismes Marins (BIOM), Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [29]Sorbonne Universités Paris 06, Observatoire Océanologique de Banyuls (OOB) UPMC, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [30]Aix Marseille Université, CNRS IGS UMR 7256, 13288 Marseille, France. [31]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan. [32]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany. [33]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME 04544, USA. [34]National Science Foundation, Arlington, VA 22230, USA. [35]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France. [36]Laboratoire de Physique des Océans, Université de Bretagne Occidentale (UBO)–Institut Universitaire Européen de la Mer (IUEM), Place Copernic, 29820 Plouzané, France.
***Corresponding author. E-mail: vargas@sb-roscoff.fr (C.d.V.); pwincker@genoscope.cns.fr (P.W.); karsenti@embl.de (E.K.)**
†These authors contributed equally to this work. ‡*Tara* Oceans Coordinators and affiliations appear at the end of this paper.

hereafter denoted "metabarcodes." We then clustered metabarcodes into biologically meaningful operational taxonomic units (OTUs) (*14*) and assigned a eukaryotic taxonomic path to all metabarcodes and OTUs by global similarity analysis with 77,449 reference, Sanger-sequenced V9 rDNA barcodes covering the known diversity of eukaryotes and assembled into an in-house database called *V9_PR2* (*15*). Beyond taxonomic assignation, we inferred basic trophic and symbiotic ecological modes (photo- versus heterotrophy; parasitism, commensalism, mutualism for both hosts and symbionts) to *Tara* Oceans reads and OTUs on the basis of their genetic affiliation to large monophyletic and monofunctional groups of reference barcodes. We finally inferred large-scale ecological patterns of eukaryotic biodiversity across geography, taxonomy, and organismal size fractions based on rDNA abundance data and community similarity analyses and compared them to current knowledge extracted from the literature.

The extent of eukaryotic plankton diversity in the photic zone of the world ocean

Sequencing of ~1.7 million V9 rDNA reads from each of the 334 size-fractionated plankton samples was sufficient to approach saturation of eukaryotic richness at both local and global scales (Fig. 1, A and B). Local richness represented, on average, 9.7 ± 4% of global richness, the latter approaching saturation at ~2 million eukaryotic metabarcodes or ~110,000 OTUs (*16*). The global pool of OTUs displayed a good fit to the truncated Preston log-normal distribution (*17*), which, by extrapolation, suggests a total photic-zone eukaryotic plankton richness of ~150,000 OTUs, of which ~40,000 were not found in our survey (Fig. 1C). Thus, we estimate that our survey unveiled ~75% of eukaryotic ribosomal diversity in the globally distributed water masses analyzed. The extrapolated ~150,000 total OTUs is much higher than the ~11,200 formally described species of marine eukaryotic plankton (see below) and probably represents a highly conservative, lower-boundary estimate of the true number of eukaryotic species in this biome, given the relatively limited taxonomic resolution power of the 18*S* rDNA gene. Our data indicate that eukaryotic taxonomic diversity is higher in smaller organismal size fractions, with a peak in the piconanoplankton (Fig. 1A), highlighting the richness of tiny organisms that are poorly characterized in terms of morphotaxonomy and physiology (*18*). A first-order, supergroup-level classification of all *Tara* Oceans OTUs demonstrated the prevalence (at the biome scale and across the >four orders of size magnitude sampled) of protist rDNA biodiversity with respect to that of classical multicellular eukaryotes, i.e., animals, plants, and fungi (Fig. 2A). Protists accounted for >85% of total eukaryotic ribosomal diversity, a ratio that may well hold true for other marine, freshwater, and terrestrial oxygenic ecosystems (*19*). The latest estimates of total marine eukaryotic biodiversity based on statistical extrapolations from classical taxonomic knowledge predict the existence of 0.5 to 2.2 million species [including all benthic and planktonic systems from reefs to deep-sea vents (*20*, *21*)] but do not take into account the protistan knowledge gap highlighted here. Simple application of our animal–to–other eukaryotes ratio of ~13% to the robust prediction of the total number of metazoan species from (*20*) would imply that 16.5 million and 60 million eukaryotic species potentially inhabit the oceans and Earth, respectively.

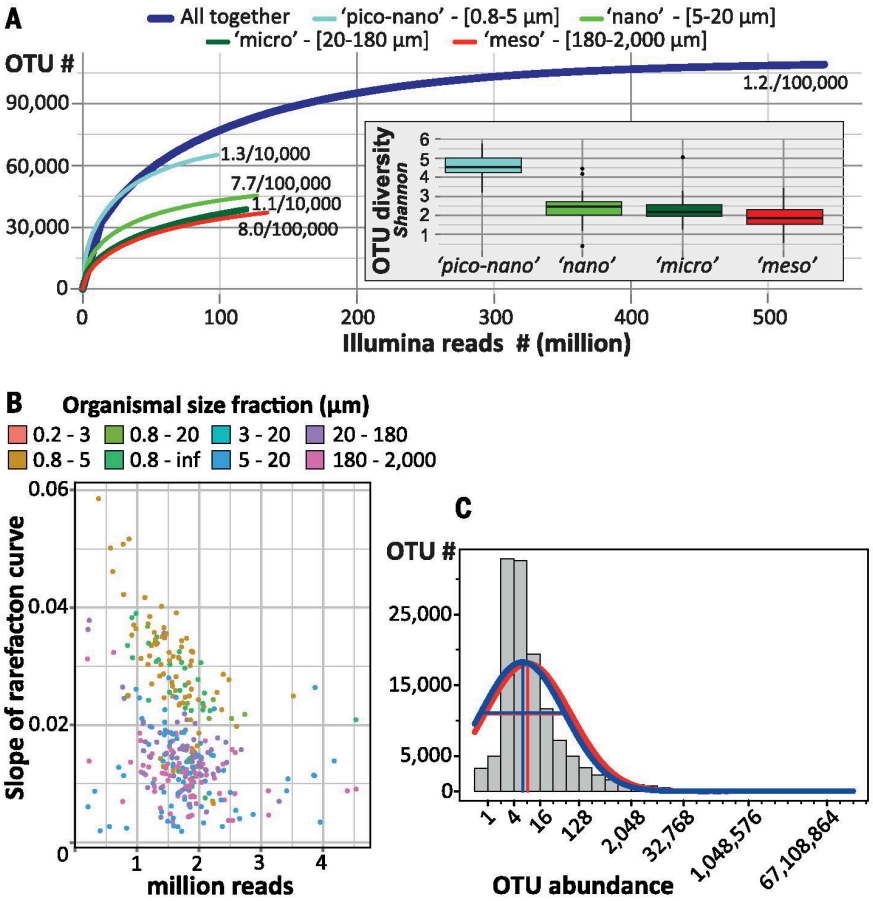


Fig. 1. Photic-zone eukaryotic plankton ribosomal diversity. (**A**) V9 rDNA OTUs rarefaction curves and overall diversity (Shannon index, inset) for each plankton organismal size fraction. Proximity to saturation is indicated by weak slopes at the end of each rarefaction curve (e.g., 1.2/100,000 means 1.2 novel metabarcodes obtained every 100,000 rDNA reads sequenced). (**B**) Saturation slope versus number of V9 rDNA reads for all of the 334 samples (dots) analyzed herein. A slope of 0.02 indicates that two novel barcodes can be recovered if 100 new reads are sequenced. Samples are colored according to size fraction. (**C**) Global OTU abundance distribution and fit to the Preston log-normal model. Most OTUs in our data set were represented by 3 to 16 reads, whereas fewer OTUs presented less or more abundances. Quasi-Poisson fit to octaves (red curve) and maximized likelihood to $\log_2$ abundances (blue curve) approximations were used to fit the OTU abundance distribution to the Preston log-normal model. Overall, the global (A) and local (B) saturation values indicate that our extensive sampling effort (in terms of spatiotemporal coverage and sequencing depth) uncovered the majority of eukaryotic ribosomal diversity within the photic layer of the world's tropical to temperate oceans. Calculation of the Preston veil, which infers the number of OTUs that we missed (or were veiled) during our sampling (~40,000), confirmed that we captured most of the protistan richness, thus allowing extraction of holistic and general patterns of eukaryotic plankton biodiversity from our data set.

Phylogenetic breakdown of photic-zone eukaryotic biodiversity

About one-third of eukaryotic ribosomal diversity in our data set did not match any reference barcode in the extensive *V9_PR2* database ("unassigned" category in Fig. 2A). This unassignable diversity represented only a small proportion (2.6%) of total reads and increased in both richness and abundance in smaller organismal size fractions, suggesting that it corresponds mostly to rare and minute taxa that have escaped previous characterization. Some may also correspond to divergent rDNA pseudogenes, known to exist in eukaryotes (*22*, *23*) or sequencing artefacts (*24*), although both of these would be expected to be present in equal proportion in all

size fractions [details in (*16*)]. The remaining ~87,000 assignable OTUs were classified into 97 deep-branching lineages covering the full spectrum of cataloged eukaryotic diversity amongst the seven recognized supergroups and multiple lineages of uncertain placement (*15*) whose origins go back to the primary radiation of eukaryotic life in the Neoproterozoic. Although highly represented in the *V9_PR2* reference database, several well-known lineages adapted to terrestrial, marine benthic, or anaerobic habitats (e.g., Embryophyta; apicomplexan and trypanosome parasites of land plants and animals; amoeboflagellate Breviatea; and several lineages of Amoebozoa, Excavata, and Cercozoa) were not detected in our metabarcoding data set, suggesting the absence of contamination during the PCR and sequencing steps on land and reducing the number of deep branches of eukaryotic plankton to 85 (Fig. 3).

We then extracted the metabarcodes assigned to morphologically well-known planktonic eukaryotic taxa from our data set and compared them with the conventional, 150 year-old morphological view of marine eukaryotic plankton that includes ~11,200 cataloged species divided into three broad categories: ~4350 species of phytoplankton (microalgae), ~1350 species of protozooplankton (relatively large, often biomineralized, heterotrophic protists), and ~5500 species of metazooplankton (holoplanktonic animals) (*25*–*27*). A congruent picture of the distribution of morphogenetic diversity among and within these organismal categories emerged from our data set (Fig. 2B), but typically, three to eight times more rDNA OTUs were found than described morphospecies in the best-known lineages within these categories. This is within the range of the number of cryptic species typically detected in globally-distributed pelagic taxa using molecular data (*28*, *29*). The general congruency between genetic and morphological data in the cataloged compartment of eukaryotic plankton suggests that the protocols used, from plankton sampling to DNA sequencing, recovered the known eukaryotic biodiversity without major qualitative or quantitative biases. However, OTUs related to morphologically described taxa represented only a minor part of the total eukaryotic plankton ribosomal and phylogenetic diversity. Overall, <1% of OTUs were strictly identical to reference sequences, and OTUs were, on average, only ~86% similar to any V9 reference sequence (Fig. 3F) (*16*). This shows that most photic-zone eukaryotic plankton V9 rDNA diversity had not been previously sequenced from cultured strains, single-cell isolates, or even environmental clone library surveys. The *Tara* Oceans metabarcode data set added considerable phylogenetic information to previous protistan rDNA knowledge, with an estimated mean tree-length increase of 453%, reaching >100% in 43 lineages (*16*). Even in the best-referenced groups such as the diatoms (1232 reference sequences) (Fig. 3B), we identified many new rDNA sequences, both within known groups and forming new clades (*16*).

Eleven "hyperdiverse" lineages each contained >1000 OTUs, together representing ~88 and

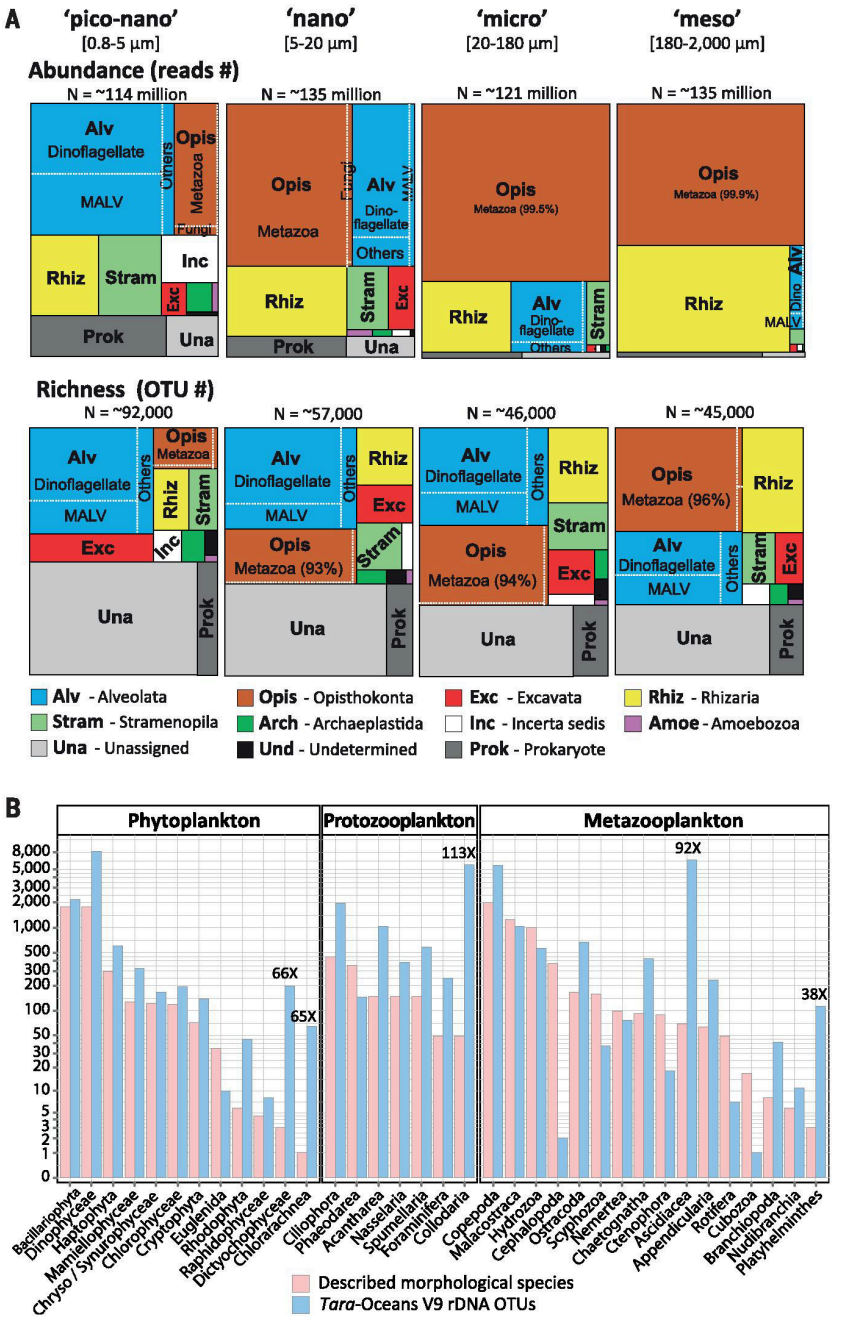


Fig. 2. Unknown and known components of eukaryotic plankton biodiversity. (**A**) Phylogenetic breakdown of the entire metabarcoding data set at the eukaryotic supergroup level. All *Tara* Oceans V9 rDNA reads and OTUs were classified among the seven recognized eukaryotic supergroups plus the known but unclassified deep-branching lineages (incertae sedis). The tree maps display the relative abundance (upper part) and richness (lower part) of the different eukaryotic supergroups in each organismal size fraction. Note that ~5% of barcodes were assigned to prokaryotes, essentially in the piconano fraction, witnessing the universality of the eukaryotic primers used. Barcodes are "unassigned" when sequence similarity to a reference sequence is <80% and "undetermined" when eukaryotic supergroups could not be discriminated (at similarity >80%). (**B**) Ribosomal DNA diversity associated with the morphologically known and cataloged part of eukaryotic plankton. The total number of morphologically described species in the literature [red bars, based on (*25*–*27*)] and the corresponding total number of *Tara* Oceans V9 rDNA OTUs (blue bars) are indicated for each of the 35 classical lineages of eukaryotic phyto-, protozoo-, and metazooplankton. The five classical groups that were found to be substantially more diverse than previously thought (from 38- to 113-fold more OTUs than morphospecies) are highlighted. Note that in the classical morphological view, phyto- and metazooplankton comprise ~88% of total eukaryotic plankton diversity.

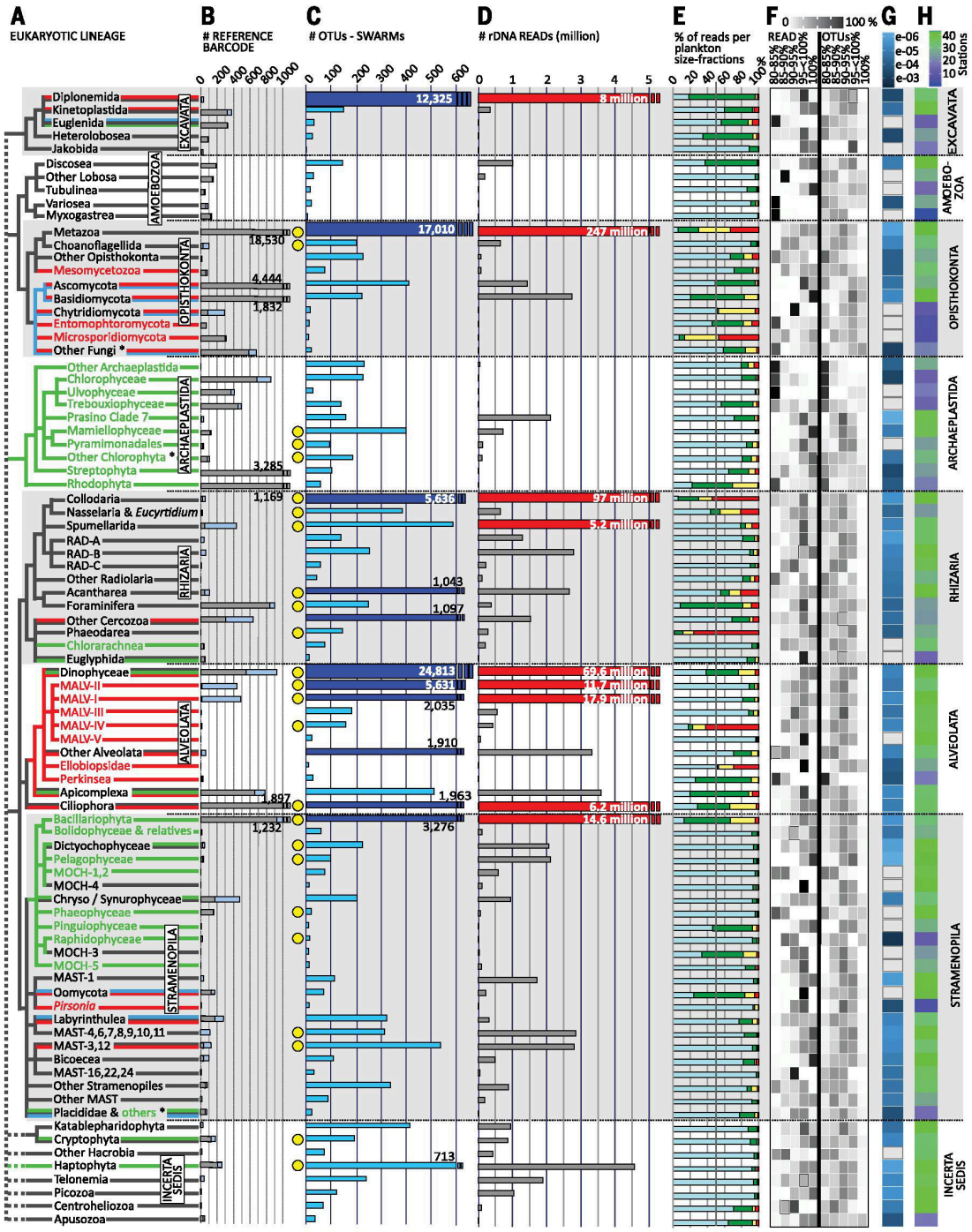


Fig. 3. Phylogenetic distribution of the assignable component of eukaryotic plankton ribosomal diversity. (**A**) Schematic phylogeny of the 85 deep-branching eukaryotic lineages represented in our global oceans metabarcoding data set, with broad ecological traits based on current knowledge: red, parasitic; green, photoautotrophic; blue, osmo- or saprotrophic; black, mostly phagotrophic lineages. Lineages known only from environmental sequence data were colored in black by default. For simplicity, three branches (denoted by asterisks) artificially group a few distinct lineages [details in (*15*)]. (**B**) Number of reference V9 rDNA barcodes used to annotate the metabarcoding data set (gray, with known taxonomy at the genus and/or species level; light blue, from previous 18S rDNA environmental clone libraries). (**C**) *Tara* Oceans V9 rDNA OTU richness. Dark blue thicker bars indicate the 11 hyperdiverse lineages containing >1000 OTUs. Yellow circles highlight the 25 lineages that have been recognized as important in previous marine plankton biodiversity and ecology studies using morphological and/or molecular data [see also (*15*)]. (**D**) Eukaryotic plankton abundance expressed as numbers of rDNA reads (the red bars indicate the nine most abundant lineages with >5 million reads). (**E**) Proportion of rDNA reads per organismal size fraction. Light blue, piconano-; green, nano-; yellow, micro-; red, mesoplankton. (**F**) Percentage of reads and OTUs with 80 to 85%, 85 to 90%, 90 to 95%, 95 to <100%, and 100% sequence similarity to a reference sequence. (**G**) Slope of OTU rarefaction curves. (**H**) Mean geographic occupancy (average number of stations in which OTUs were observed, weighted by OTU abundance).

~90% of all OTUs and reads, respectively (Fig. 3C). Among these, the only permanently phototrophic taxa were diatoms (Fig. 4A) and about one-third of dinoflagellates (Fig. 4, B to F), together comprising ~15 and ~13% of hyperdiverse OTUs and reads, respectively (*30*). Most hyperdiverse photic-zone plankton belonged to three supergroups—the Alveolata, Rhizaria, and Excavata—about which we have limited biological or ecological information. The Alveolata, which consist mostly of parasitic [marine alveolates (MALVs)] (Fig. 4F) and phagotrophic (ciliates and most dinoflagellates) taxa, were by far the most diverse supergroup, comprising ~42% of all assignable OTUs. The Rhizaria are a group of amoeboid heterotrophic protists with active pseudopods displaying a broad spectrum of ecological behavior, from phagotrophy to parasitism and mutualism (symbioses) (*31*). Rhizarian diversity peaked in the Retaria (Fig. 4, C and D) a subgroup including giant protists that build complex skeletons of silicate (Polycystinea), strontium sulfate (Acantharia) (Fig. 4C), or calcium carbonate (Foraminifera) and thus comprise key microfossils for paleoceanography. Unsuspected rDNA diversity was recorded within the Collodaria (5636 OTUs), polycystines that are mostly colonial, poorly silicified, or naked and live in obligatory symbiosis with photosynthetic dinoflagellates (Fig. 4D) (*32*, *33*). Arguably, the most surprising component of novel biodiversity was the >12,300 OTUs related to reference sequences of diplonemids, an excavate lineage that has only two described genera of flagellate grazers, one of which parasitizes diatoms and crustaceans (*34*, *35*). Their ribosomal diversity was not only much higher than that observed in classical plankton groups such as foraminifers, ciliates, or diatoms (50-fold, 6-fold, and 3.8-fold higher, respectively) but was also far from richness saturation (Fig. 3E). Eukaryotic rDNA diversity peaked especially in the few lineages that extend across larger size fractions (i.e., metazoans, rhizarians, dinoflagellates, ciliates, diatoms) (Fig. 3E). Larger cells or colonies not only provide protection against predation via size-mediated avoidance and/or construction of composite skeletons but also provide support for complex and coevolving relationships with often specialized parasites or mutualistic symbionts.

Beyond this hyperdiverse, largely heterotrophic eukaryotic majority, our data set also highlighted the phylogenetic diversity of poorly known phagotrophic (e.g., 413 OTUs of Katablepharidophyta, 240 OTUs of Telonemia), osmotrophic (e.g., 410 OTUs of Ascomycota, 322 OTUs of Labyrinthulea), and parasitic (e.g., 384 OTUs of gregarine apicomplexans, 160 OTUs of Ascetosporea, 68

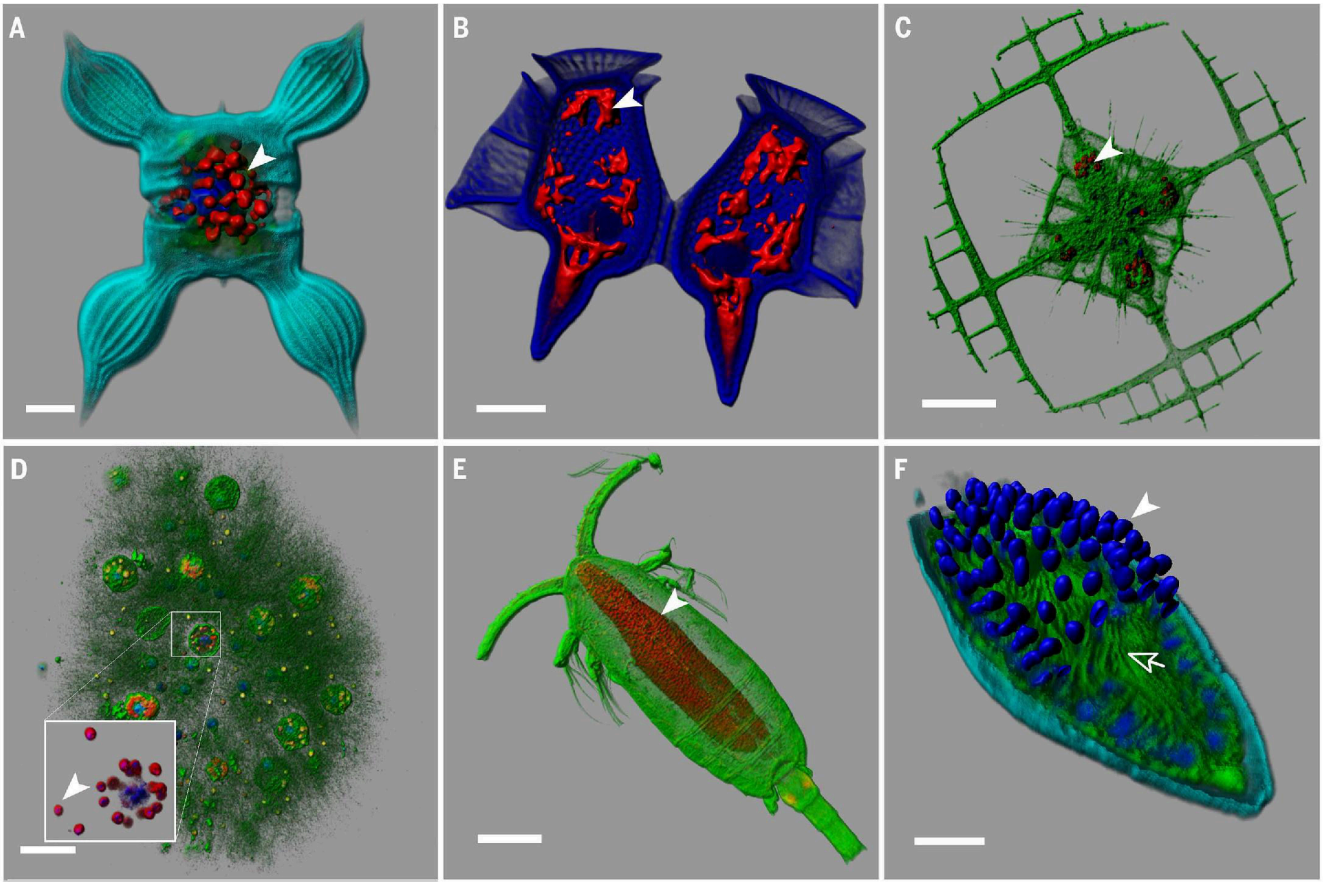


Fig. 4. Illustration of key eukaryotic plankton lineages. (**A**) Stramenopila; a phototrophic diatom *Chaetoceros bulbosus*, with its chloroplasts in red (arrowhead). Scale bar, 10 μm. (**B**) Alveolata; a heterotrophic dinoflagellate *Dinophysis caudata* harboring kleptoplasts [in red (arrowhead)]. Scale bar, 20 μm (*75*). (**C**) Rhizaria; an acantharian *Lithoptera* sp. with endosymbiotic haptophyte cells from the genus *Phaeocystis* [in red (arrowhead)]. Scale bar, 50 μm (*41*). (**D**) Rhizaria; inside a colonial network of Collodaria, a cell surrounded by several captive dinoflagellate symbionts of the genus *Brandtodinium* (arrowhead). Scale bar, 50 μm (*33*). (**E**) Opisthokonta; a copepod whose gut is colonized by the parasitic dinoflagellate *Blastodinium* [red area shows nuclei (arrowhead)]. Scale bar, 100 μm (*51*). (**F**) Alveolata; a cross-sectioned, dinoflagellate cell infected by the parasitoid alveolate *Amoebophrya* (MALV-II). Each blue spot (arrowhead) is the nucleus of future free-living dinospores; their flagella are visible in green inside the mastigocoel cavity (arrow). Scale bar, 5 μm. The cellular membranes were stained with DiOC6 (green); DNA and nuclei were stained with Hoechst (blue) [the dinoflagellate theca in (B) was also stained by this dye]. Chlorophyll autofluorescence is shown in red [except for in (E)]. An unspecific fluorescent painting of the cell surface (light blue) was used to reveal cell shape for (A) and (F). All specimens come from *Tara* Oceans samples preserved for confocal laser scanning fluorescent microscopy. Images were three-dimensionally reconstructed with Imaris (Bitplane).

OTUs of Ichthyosporea) protist groups. Amongst the 85 major lineages presented in the phylogenetic framework of Fig. 3, less than one-third (~25) have been recognized as important in previous marine plankton biodiversity and ecology studies using morphological and/or molecular data (Fig. 3C) (*15*). The remaining ~60 branches had either never been observed in marine plankton or were detected through morphological description of one or a few species and/or the presence of environmental sequences in geographically restricted clone library surveys (*15*). This understudied diversity represents ~25% of all taxonomically assignable OTUs (>21,500) and covers broad taxonomic and geographic scales, thus representing a wealth of new actors to integrate into future plankton systems biology studies.

Insights into photic-zone eukaryotic plankton ecology

Functional annotation of taxonomically assigned V9 rDNA metabarcodes was used as a first attempt to explore ecological patterns of eukaryotic diversity across broad spatial scales and organismal size fractions, focusing on fundamental trophic modes (photo- versus heterotrophy) and symbiotic interactions (parasitism to mutualism). Heterotroph (protists and metazoans) V9 rDNA metabarcodes were substantially more diverse (63%) and abundant (62%) than phototroph metabarcodes that represented <20% of OTUs and reads across all size fractions and geographic sites, with an increasing heterotroph-to-phototroph ratio in the micro- and mesoplankton (Fig. 5A, confirmed in 17 non–size-fractionated samples (*30*). These results challenge the classical morphological view of plankton diversity, biased by a terrestrial ecology approach, whereby phyto- and metazooplankton (the plant-animal paradigm) are thought to comprise ~88% of eukaryotic plankton diversity (Fig. 2B) and heterotrophic protists are typically reduced in food-web modeling to a single entity, often idealized as ciliate grazers.

An unsuspected richness and abundance of metabarcodes assigned to monophyletic groups of heterotrophic protists that cannot survive without endosymbiotic microalgae was found in larger size fractions ("photosymbiotic hosts" in Fig. 5A). Their abundance and even diversity were sometimes greater than those of all metazoan metabarcodes, including those from copepods. Most of these cosmopolitan photosymbiotic hosts were found within the hyperdiverse radiolarians Acantharia (1043 OTUs) and Collodaria (5636 OTUs) (Figs. 3, 4B, and 5D), which have often been overlooked in traditional morphological surveys of plankton-net–collected material because of their delicate gelatinous and/or easily dissolved structures but are known to be very abundant from microscope-based and in situ imaging studies (*36*–*38*). All 95 known colonial collodarian species described since the 19th century (*39*) harbor intracellular symbiotic microalgae, and these key players for plankton ecology are protistan analogs of photosymbiotic corals in tropical coastal reef ecosystems with no equivalent in terrestrial ecology. In addition to their contribution to total primary production (*36*, *38*), these diverse, biologically complex, often biomineralized, and relatively long-lived giant mixotrophic protists stabilize carbon in larger size fractions and probably increase its flux to the ocean interior (*38*). Conversely, the microalgae that are known obligate intracellular partners in open-ocean photosymbioses (*33*, *40*–*42*) (Fig. 5B) were neither very diverse nor highly abundant and occurred evenly across organismal size fractions (Fig. 5C). However, their relative contribution was greatest in the mesoplankton category (10%) (Fig. 5C), where the known photosymbionts of pelagic rhizarians were found (together with their hosts) (Fig. 5B). The stable and systematic abundance of photosymbiotic microalgae across size fractions [a pattern not shown by nonphotosymbiotic microalgae (*30*)] suggests that pelagic photosymbionts maintain free-living and potentially actively growing populations in the piconano- and nanoplankton, representing an accessible pool for recruitment by their heterotrophic hosts. This appears to contrast with photosymbioses in coral reefs and terrestrial systems, where symbiotic microalgal populations mainly occur within their multicellular hosts (*43*).

On the other end of the spectrum of biological interactions, rDNA metabarcodes affiliated to groups of known parasites were ~90 times more diverse than photosymbionts in the piconanoplankton, where they represented ~59% of total heterotrophic protistan ribosomal richness and ~53% of abundance (Figs. 4 and 5C), although this latter value may be inflated by a hypothetically higher rDNA copy number in some marine alveolate lineages (*18*). Parasites in this size fraction were mostly (89% of diversity and 88% of abundance across all stations) within the MALV-I and -II Syndiniales (*30*), which are known exclusively as parasitoid species that kill their hosts and release hundreds of small (2 to 10 μm), nonphagotrophic dinospores (*9*, *44*) that survive for only a few days in the water column (*45*). Abundant parasite-assigned metabarcodes in small size fractions (Fig. 5, B and C) suggest the existence of a large and diverse pool of free-living parasites in photic-zone piconanoplankton, mirroring phage ecology (*46*) and reflecting the extreme diversity and abundance of their known main hosts: radiolarians, ciliates, and dinoflagellates (Fig. 3) (*9*, *47*–*49*). Contrasting with the pattern observed for metabarcodes affiliated to purely phagotrophic taxa, the relative abundance and richness of putative parasite metabarcodes decreased in the nano- and microplanktonic size fractions but increased again in the mesoplankton (Fig. 5C), where parasites are most likely in their infectious stage within larger-sized host organisms. This putative in hospite parasites richness, equivalent to only 23% of that in the piconanoplankton, consisted mostly of a variety of alveolate taxa known to infect crustaceans: MALV-IV such as *Haematodinium* and *Syndinium*; dinoflagellates such as *Blastodinium* (Fig. 4E); and apicomplexan gregarines, mainly *Cephaloidophoroidea* (Fig. 5B) (*9*, *50*, *51*). This pattern contrasts with terrestrial systems where most parasites live within their hosts and are typically transmitted either vertically or through vectors because they generally do not survive outside their hosts (*52*). In the pelagic realm, free-living parasitic spores, like phages, are protected from dessication and dispersed by water diffusion and are apparently massively produced, which likely increases horizontal transmission rate.

Community structuring of photic-zone eukaryotic plankton

Clustering of communities by their compositional similarity revealed the primary influence of organism size ($P = 10^{-3}$, $r^2 = 0.73$) on community structuring, with piconanoplankton displaying stronger cohesiveness than larger organismal size fractions (Fig. 6A). Filtered size-fraction–specific communities separated by thousands of kilometers were more similar in composition than they were to communities from other size fractions at the same location. This was emphasized by the fact that ~36% of all OTUs were restricted to a single size category (*53*). Further analyses within each organismal size fraction indicated that geography plays a role in community structuring, with samples being partially structured according to basin of origin, a pattern that was stronger in larger organismal size fractions ($P = 0.001$ in all cases, $r^2 = 0.255$ for piconanoplankton, 0.371 for nanoplankton, 0.473 for microplankton, and 0.570 for mesoplankton) (Fig. 6B). Mantel correlograms comparing Bray-Curtis community similarity to geographic distances between all samples indicated significant positive correlations in all organismal size fractions over the first ~6000 km, the correlation breaking down at larger geographic distances (*54*). This positive correlation between community dissimilarity and geographic distance, expected under neutral biodiversity dynamics (*55*), challenges the classical niche model for photic-zone eukaryotic plankton biogeography (*56*). The significantly stronger community differentiation by ocean basin in larger organismal size fractions (Fig. 6B) suggests increasing dispersal limitation from piconano- to nano-, micro-, and mesoplankton. Thus, larger-sized eukaryotic plankton communities, containing the highest abundance and diversity of metazoans (Figs. 2A and 5B), were spatially more heterogeneous in terms of both taxonomic (Fig. 6) and functional (Fig. 5A) composition and abundance. The complex life cycle and behaviors of metazooplankton, including temporal reproductive and growth cycles and vertical migrations, together with putative rapid adaptive evolution processes to mesoscale oceanographic features (*57*), may explain the stronger geographic differentiation of mesoplanktonic communities. By contrast, eukaryotic communities in the piconanoplankton were richer (Fig. 1A) and more homogeneous in taxonomic composition (Fig. 6), representing a stable compartment across the world's oceans (*58*).

Even though protistan communities were diverse, the proportions of abundant (>1%) and

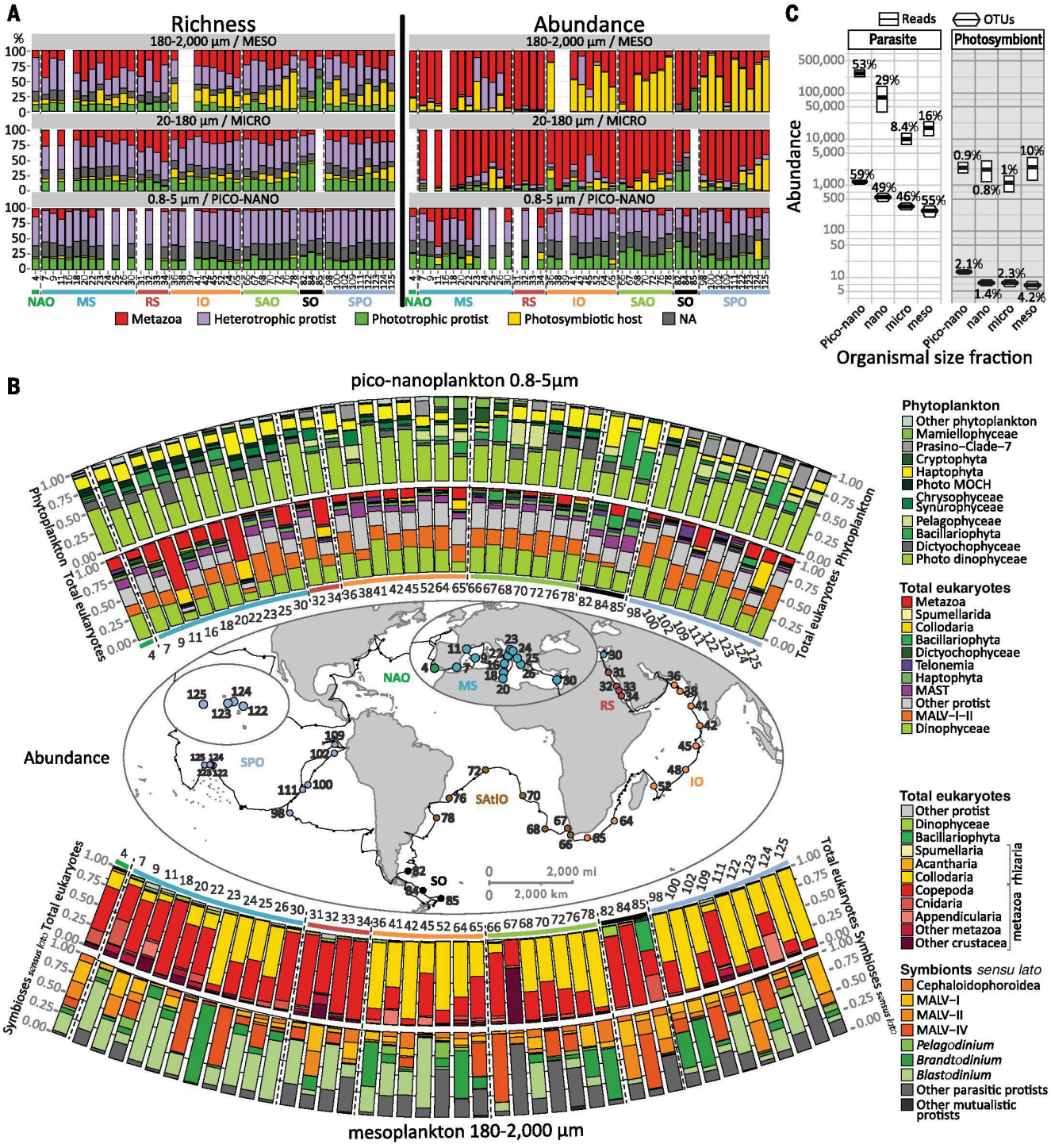


Fig. 5. Metabarcoding inference of trophic and symbiotic ecological diversity of photic-zone eukaryotic plankton. (**A**) Richness (OTU number) and abundance (read number) of rDNA metabarcodes assigned to various trophic taxo-groups across plankton organismal size fractions and stations. Note that the nano size fraction did not contain enough data to be used in this biogeographical analysis [for all size-fraction data, see (*30*)]. NA, not applicable. (**B**) Relative abundance of major eukaryotic taxa across *Tara* Oceans stations for (i) phytoplankton and all eukaryotes in piconanoplankton (above the map) and (ii) all eukaryotes and protistan symbionts (sensu lato) in mesoplankton (below the map). Note the pattern of inverted relative abundance between collodarian colonies (Fig. 4) and copepods in, respectively, the oligotrophic and eutrophic and mesotrophic systems. The dinoflagellates *Brandtodinium* and *Pelagodinium* are endophotosymbionts in Collodaria (*33*) and Foraminifera (*40*, *42*), respectively. (**C**) Richness and abundance of parasitic and photosymbiotic (microalgae) protists across organismal size fractions. The relative contributions (percent) of parasites to total heterotrophic protists and of photosymbionts to total phytoplankton are indicated above each symbol.

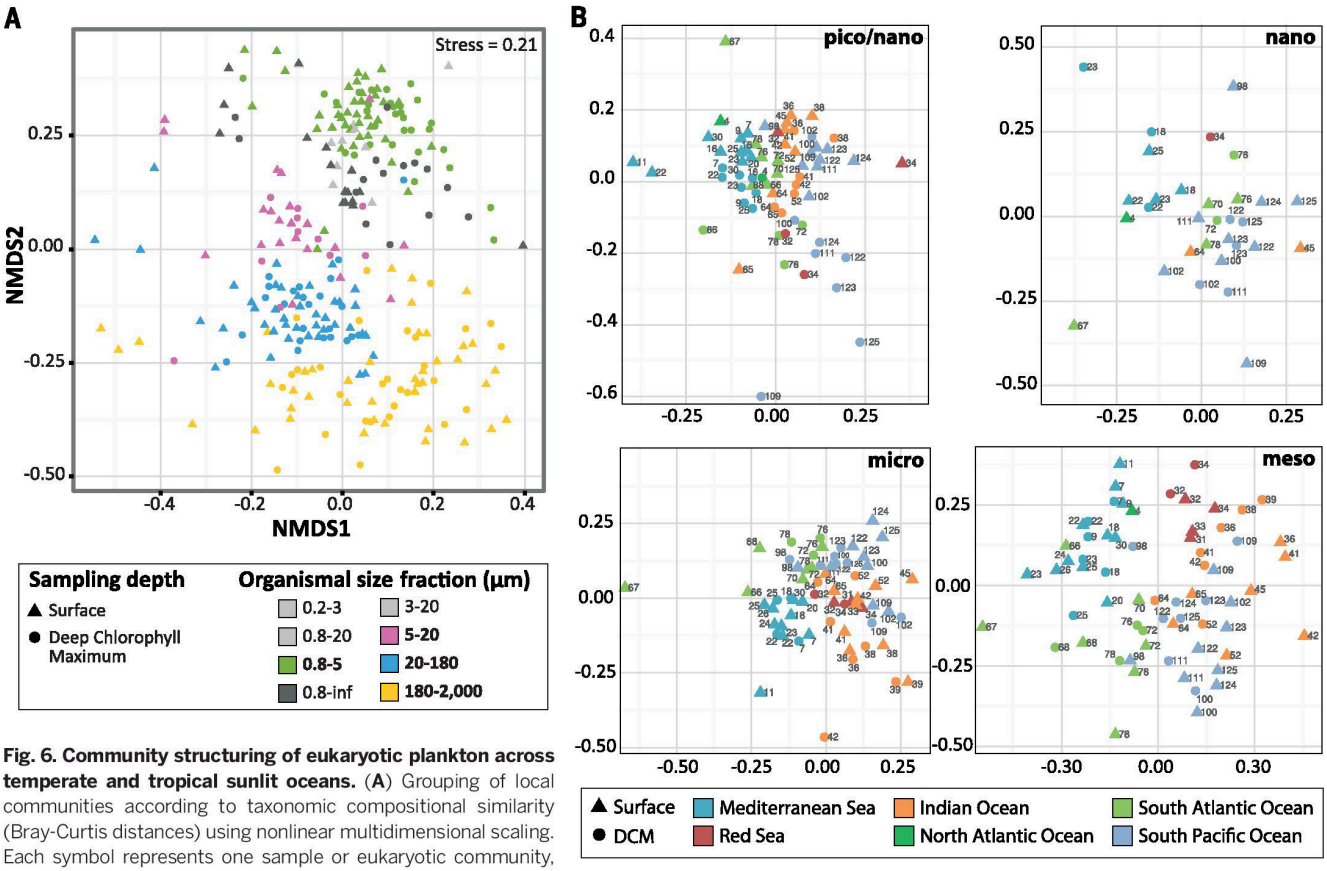


Fig. 6. Community structuring of eukaryotic plankton across temperate and tropical sunlit oceans. (**A**) Grouping of local communities according to taxonomic compositional similarity (Bray-Curtis distances) using nonlinear multidimensional scaling. Each symbol represents one sample or eukaryotic community, corresponding to a particular depth (shape) and organismal size fraction (color). (**B**) Same as in (A), but the different plankton organismal size fractions were analyzed independently, and communities are distinguished by depth (shape) and ocean basins' origin (color). An increasing geographic community differentiation along increasing organismal size fractions is visible and confirmed by the Mantel test [$P = 10^{-3}$, R_m = 0.36, 0.49, 0.50, and 0.51 for the highest piconano- to mesoplankton correlations in Mantel correlograms; see also (*54*)]. In addition, samples from the piconanoplankton only were discriminated by depth (surface versus DCM; $P = 0.001$, $r^2 = 0.2$). The higher diversity and abundance of eukaryotic phototrophs in this fraction (Fig. 5A) may explain overall community structuring by light and, thus, depth.

rare (<0.01%) OTUs were more or less constant across communities, as has been observed in coastal waters (*6*). Only 2 to 17 OTUs (i.e., 0.2 to 8% of total OTUs per and across sample) dominated each community (*54*), suggesting that a small proportion of eukaryotic taxa are key for local plankton ecosystem function. On a worldwide scale, an occurrence-versus-abundance analysis of all ~110,000 *Tara* Oceans OTUs revealed the hyperdominance of cosmopolitan taxa (Fig. 7A). The 381 (0.35% of the total) cosmopolitan OTUs represented ~68% of the total number of reads in the data set. Of these, 269 (71%) OTUs had >100,000 reads and accounted for nearly half (48%) of all rDNA reads (Fig. 7A), a pattern reminiscent of hyperdominance in the largest forest ecosystem on Earth, where only 227 tree species out of an estimated total of 16,000 account for half of all trees in Amazonia (*59*). The cosmopolitan OTUs belonged mainly (314 of 381) to the 11 hyperdiverse eukaryotic planktonic lineages (Fig. 3C) and were essentially phagotrophic (40%) or parasitic (21%), with relatively few (15%) phytoplanktonic taxa (*54*). Of the cosmopolitan OTUs, which represent organisms that are likely among the most abundant eukaryotes on Earth, 25% had poor identity (<95%) to reference taxa, and 11 of these OTUs could not even be affiliated to any available reference sequence (Fig. 7B) (*54*).

Conclusions and perspectives

We used rDNA sequence data to explore the taxonomic and ecological structure of total eukaryotic plankton from the photic oceanic biome, and we integrated these data with existing morphological knowledge. We found that eukaryotic plankton are more diverse than previously thought, especially heterotrophic protists, which may display a wide range of trophic modes (*60*) and include an unsuspected diversity of parasites and photosymbiotic taxa. Dominance of unicellular heterotrophs in plankton ecosystems likely emerged at the dawn of the radiation of eukaryotic cells, together with arguably their most important innovation: phagocytosis. The onset of eukaryophagy in the Neoproterozoic (*61*) probably led to adaptive radiation in heterotrophic eukaryotes through specialization of trophic modes and symbioses, opening novel serial biotic ecological niches. The extensive codiversification of relatively large heterotrophic eukaryotes and their associated parasites supports the idea that biotic interactions, rather than competition for resources and space (*62*), are the primary forces driving organismal diversification in marine plankton systems. Based on rDNA, heterotrophic protists may be even more diverse than prokaryotes in the planktonic ecosystem (*63*). Given that organisms in highly diverse and abundant groups, such as the alveolates and rhizarians, can have genomes more complex than those of humans (*64*), eukaryotic plankton may contain a vast reservoir of unknown marine planktonic genes (*65*). Insights are developing into how heterotrophic protists contribute to a multilayered and integrated ecosystem. The protistan parasites and mutualistic symbionts increase connectivity and complexity of pelagic food webs (*66*, *67*) while contributing to the carbon quota of their larger, longer-lived, and often biomineralized symbiotic hosts, which themselves contribute to carbon export when they die. Decoding the ecological and evolutionary rules governing plankton diversity remains essential for understanding how the

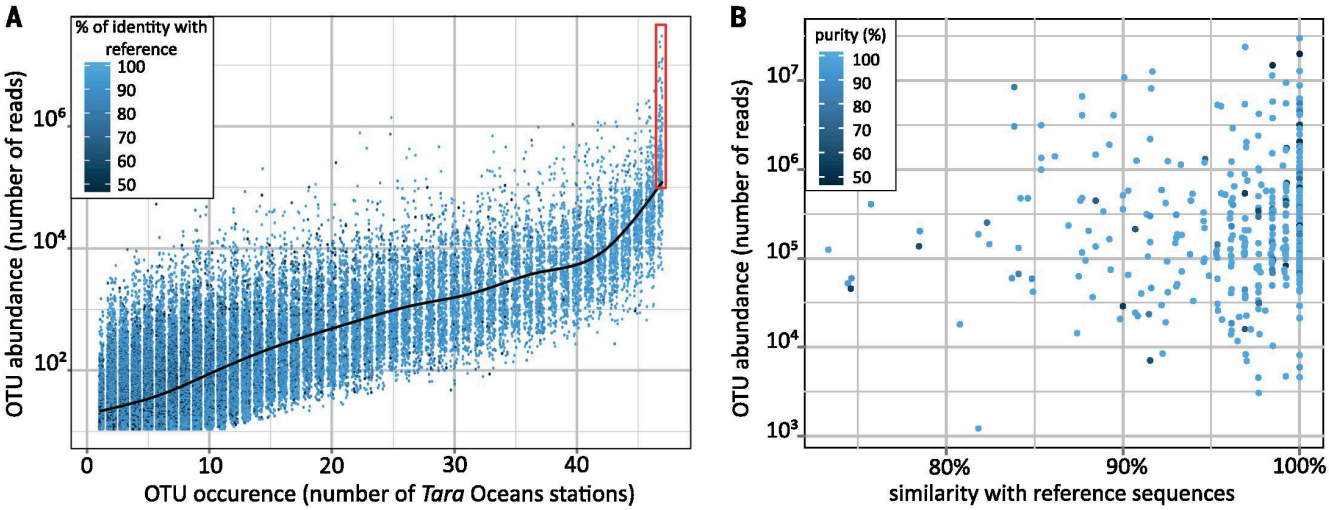


Fig. 7. Cosmopolitanism and abundance of eukaryotic marine plankton. (**A**) Occurrence-versus-abundance plot including the ~110,000 *Tara* Oceans V9 rDNA OTUs. OTUs are colored according to their identity with a reference sequence, and a fitted curve indicates the median OTU size value for each OTU geographic occurrence value. The red rectangle encloses the cosmopolitan and hyperdominant (>10^5 reads) OTUs. (**B**) Similarity to reference barcode and taxonomic purity [a measure of taxonomic assignment consistency defined as the percentage of reads within an OTU assigned to the same taxon; see (*13*)] of the 381 cosmopolitan OTUs, along their abundance (*y* axis).

critical ocean biomes contribute to the functioning of the Earth system.

Materials and methods

V9-18S rDNA for eukaryotic metabarcoding

We used universal eukaryotic primers (*68*) to PCR-amplify (25 cycles in triplicate) the V9-18*S* rDNA genes from all *Tara* Oceans samples. This barcode presents a combination of advantages for addressing general questions of eukaryotic biodiversity over extensive taxonomic and ecological scales: (i) It is universally conserved in length (130 ± 4 base pairs) and simple in secondary structure, thus allowing relatively unbiased PCR amplification across eukaryotic lineages followed by Illumina sequencing. (ii) It includes both stable and highly variable nucleotide positions over evolutionary time frames, allowing discrimination of taxa over a substantial phylogenetic depth. (iii) It is extensively represented in public reference databases across the eukaryotic tree of life, allowing taxonomic assignment among all known eukaryotic lineages (*13*).

Biodiversity analyses

Our bioinformatic pipeline included quality checking (Phred score filtering, elimination of reads without perfect forward and reverse primers, and chimera removal) and conservative filtering (removal of metabarcodes present in less than three reads and two distinct samples). The ~2.3 million metabarcodes (distinct reads) were clustered using an agglomerative, unsupervised single-linkage clustering algorithm, allowing OTUs to reach their natural limits while avoiding arbitrary global clustering thresholds (*13*, *14*). This clustering limited overestimation of biodiversity due to errors in PCR amplification or DNA sequencing, as well as intragenomic polymorphism of rDNA gene copies (*13*). *Tara* Oceans metabarcodes and OTUs were taxonomically assigned by comparison to the 77,449 reference barcodes included in our *V9_PR2* database (*15*). This database derives from the Protist Ribosomal Reference (PR2) database (*69*) but focuses on the V9 region of the gene and includes the following reorganizations: (i) extension of the number of ranks for groups with finer taxonomy (e.g., animals), (ii) expert curation of the taxonomy and renaming in novel environmental groups and dinoflagellates, (iii) resolution of all taxonomic conflicts and inclusion of environmental sequences only if they provide additional phylogenetic information, and (iv) annotation of basic trophic and/or symbiotic modes for all reference barcodes assigned to the genus level [see (*53*) and (*15*) for details]. The *V9_PR2* reference barcodes represent 24,435 species and 13,432 genera from all known major lineages of the tree of eukaryotic life (*15*). Metabarcodes with ≥80% identity to a reference V9 rDNA barcode were considered assignable. Below this threshold it is not possible to discriminate between eukaryotic supergroups, given the short length of V9 rDNA sequences and the relatively fast rate accumulation of substitution mutations in the DNA. In addition to assignment at the finest-possible taxonomic resolution, all assignable metabarcodes were classified into a reference taxonomic framework consisting of 97 major monophyletic groups comprising all known high-rank eukaryotic diversity. This framework, primarily based on a synthesis of protistan biodiversity (*19*), also included all key but still unnamed planktonic clades revealed by previous environmental rDNA clone library surveys (*70*) [e.g., marine alveolates (MALV), marine stramenopiles (MAST), marine ochrophytes (MOCH), and radiolarians (RAD)] (*15*). Details of molecular and bioinformatics methods are available on a companion Web site at http://taraoceans.sb-roscoff.fr/EukDiv/ (*53*). We compiled our data into two databases including the taxonomy, abundance, and size fraction and biogeography information associated with each metabarcode and OTU (*71*).

Ecological inferences

From our *Tara* Oceans metabarcoding data set, we inferred patterns of eukaryotic plankton functional ecology. Based on a literature survey, all reference barcodes assigned to at least the genus level that recruited *Tara* Oceans metabarcodes were associated to basic trophic and symbiotic modes of the organism they come from (*15*) and used for a taxo-functional annotation of our entire metabarcoding data set with the same set of rules used for taxonomic assignation (*53*). False positives were minimized by (i) assigning ecological modes to all individual reference barcodes in *V9_PR2*; (ii) inferring ecological modes to metabarcodes related to monomodal reference barcode(s) (otherwise transferring them to a "NA, nonapplicable" category); and (iii) exploring broad and complex trophic and symbiotic modes that involve fundamental reorganization of the cell structure and metabolism, emerged relatively rarely in the evolutionary history of eukaryotes, and most often concern all known species within monophyletic and ancient groups [see (*15*) for details]. In case of photo- versus heterotrophy, >75% of the major, deep-branching eukaryotic lineages considered (Fig. 3) are monomodal and recruit ~87 and ~69% of all *Tara* Oceans V9 rDNA reads and OTUs, respectively. For parasitism, ~91% of *Tara* Oceans metabarcodes are falling within monophyletic and major groups containing exclusively parasitic species (essentially within the major MALVs groups). Although biases could arise in functional annotation of metabarcodes

relatively distant from reference barcodes in the few complex polymodal groups (e.g., the dinoflagellates that can be phototrophic, heterotrophic, parasitic, or photosymbiotic), a conservative analysis of the trophic and symbiotic ecological patterns presented in Fig. 3, using a ≥99% assignation threshold, shows that these are stable across organismal size fractions and space, independently of the similarity cutoff (80 or 99%), demonstrating their robustness across evolutionary times (*30*).

Note that rDNA gene copy number varies from one to thousands in single eukaryotic genomes (*72*, *73*), precluding direct translation of rDNA read number into abundance of individual organisms. However, the number of rDNA copies per genome correlates positively to the size (*73*) and particularly to the biovolume (*72*) of the eukaryotic cell it represents. We compiled published data from the last ~20 years, confirming the positive correlation between eukaryotic cell size and rDNA copy number across a wide taxonomic and organismal size range [see (*74*); note, however, the ~one order of magnitude of cell size variation for a given rDNA copy number]. To verify whether our molecular ecology protocol preserved this empirical correlation, light microscopy counts of phytoplankton belonging to different eukaryotic supergroups (coccolithophores, diatoms, and dinoflagellates) were performed from nine *Tara* Oceans stations from the Indian, Atlantic, and Southern oceans; transformed into biomass and biovolume data; and then compared with the relative number of V9 rDNA reads found for the identified taxa in the same samples (*74*). Results confirmed the correlation between biovolume and V9 rDNA abundance data ($r^2 = 0.97$, $P = 1 \times 10^{-16}$), although we cannot rule out the possibility that some eukaryotic taxa may not follow the general trend.

REFERENCES AND NOTES

1. C. B. Field, M. J. Behrenfeld, J. T. Randerson, P. Falkowski, Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* **281**, 237–240 (1998). doi: 10.1126/science.281.5374.237; pmid: 9657713
2. D. A. Caron, P. D. Countway, A. C. Jones, D. Y. Kim, A. Schnetzer, Marine protistan diversity. *Annu. Rev. Mar. Sci.* **4**, 467–493 (2012). doi: 10.1146/annurev-marine-120709-142802; pmid: 22457984
3. P. López-García, F. Rodríguez-Valera, C. Pedrós-Alió, D. Moreira, Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409**, 603–607 (2001). doi: 10.1038/35054537; pmid: 11214316
4. S. Y. Moon-van der Staay, R. De Wachter, D. Vaulot, Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409**, 607–610 (2001). doi: 10.1038/35054541; pmid: 11214317
5. B. Díez, C. Pedrós-Alió, R. Massana, Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. *Appl. Environ. Microbiol.* **67**, 2932–2941 (2001). doi: 10.1128/AEM.67.7.2932-2941.2001; pmid: 11425705
6. R. Logares *et al.*, Patterns of rare and abundant marine microbial eukaryotes. *Curr. Biol.* **24**, 813–821 (2014). doi: 10.1016/j.cub.2014.02.050; pmid: 24704080
7. V. Edgcomb *et al.*, Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. *ISME J.* **5**, 1344–1356 (2011). doi: 10.1038/ismej.2011.6; pmid: 21390079
8. R. Massana *et al.*, Phylogenetic and ecological analysis of novel marine stramenopiles. *Appl. Environ. Microbiol.* **70**, 3528–3534 (2004). doi: 10.1128/AEM.70.6.3528-3534.2004; pmid: 15184153
9. L. Guillou *et al.*, Widespread occurrence and genetic diversity of marine parasitoids belonging to *Syndiniales* (*Alveolata*). *Environ. Microbiol.* **10**, 3349–3365 (2008). doi: 10.1111/j.1462-2920.2008.01731.x; pmid: 18771501
10. H. Liu *et al.*, Extreme diversity in noncalcifying haptophytes explains a major pigment paradox in open oceans. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 12803–12808 (2009). doi: 10.1073/pnas.0905841106; pmid: 19622724
11. Companion Web site: Figure W1 and Database W1 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
12. S. Pesant *et al.*, Open science resources for the discovery and analysis of *Tara* Oceans data. http://biorxiv.org/content/early/2015/05/08/019117 (2015).
13. Companion Web site: Text W1 and Figure W2 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
14. F. Mahé, T. Rognes, C. Quince, C. de Vargas, M. Dunthorn, Swarm: Robust and fast clustering method for amplicon-based studies. *PeerJ* **2**, e593 (2014). doi: 10.7717/peerj.593; pmid: 25276506
15. Companion Web site: Database W2, Database W3, and Database W6 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
16. Companion Web site: Text W3, Text W4, Text W5, Figure W4, Figure W5, Figure W6, and Figure W7 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
17. F. W. Preston, The commonness, and rarity, of species. *Ecology* **29**, 254–283 (1948). doi: 10.2307/1930989
18. R. Massana, Eukaryotic picoplankton in surface oceans. *Annu. Rev. Microbiol.* **65**, 91–110 (2011). doi: 10.1146/annurev-micro-090110-102903; pmid: 21639789
19. J. Pawlowski *et al.*, CBOL protist working group: Barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLOS Biol.* **10**, e1001419 (2012). doi: 10.1371/journal.pbio.1001419; pmid: 23139639
20. C. Mora, D. P. Tittensor, S. Adl, A. G. B. Simpson, B. Worm, How many species are there on Earth and in the ocean? *PLOS Biol.* **9**, e1001127 (2011). doi: 10.1371/journal.pbio.1001127; pmid: 21886479
21. W. Appeltans *et al.*, The magnitude of global marine species diversity. *Curr. Biol.* **22**, 2189–2202 (2012). doi: 10.1016/j.cub.2012.09.036; pmid: 23159596
22. L. M. Márquez, D. J. Miller, J. B. MacKenzie, M. J. H. Van Oppen, Pseudogenes contribute to the extreme diversity of nuclear ribosomal DNA in the hard coral *Acropora*. *Mol. Biol. Evol.* **20**, 1077–1086 (2003). doi: 10.1093/molbev/msg122; pmid: 12777522
23. S. R. Santos, R. A. Kinzie III, K. Sakai, M. A. Coffroth, Molecular characterization of nuclear small subunit (18S)-rDNA pseudogenes in a symbiotic dinoflagellate (*Symbiodinium*, Dinophyta). *J. Eukaryot. Microbiol.* **50**, 417–421 (2003). doi: 10.1111/j.1550-7408.2003.tb00264.x; pmid: 14733432
24. J. Decelle, S. Romac, E. Sasaki, F. Not, F. Mahé, Intracellular diversity of the V4 and V9 regions of the 18S rRNA in marine protists (radiolarians) assessed by high-throughput sequencing. *PLOS ONE* **9**, e104297 (2014). doi: 10.1371/journal.pone.0104297; pmid: 25090095
25. A. Sournia, M.-J. Chrétiennot-Dinet, M. Ricard, Marine phytoplankton: How many species in the world ocean? *J. Plankton Res.* **13**, 1093–1099 (1991). doi: 10.1093/plankt/13.5.1093
26. P. H. Wiebe *et al.*, Deep-sea sampling on CMarZ cruises in the Atlantic Ocean – An introduction. *Deep-Sea Res. Part II* **57**, 2157–2166 (2010). doi: 10.1016/j.dsr2.2010.09.018
27. D. Boltovskoy, Diversity and endemism in cold waters of the South Atlantic: Contrasting patterns in the plankton and the benthos. *Sci. Mar.* **69**, 17–26 (2005).
28. C. de Vargas, R. Norris, L. Zaninetti, S. W. Gibb, J. Pawlowski, Molecular evidence of cryptic speciation in planktonic foraminifers and their relation to oceanic provinces. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 2864–2868 (1999). doi: 10.1073/pnas.96.6.2864; pmid: 10077602
29. K. M. K. Halbert, E. Goetze, D. B. Carlon, High cryptic diversity across the global range of the migratory planktonic copepods *Pleuromamma piseki* and *P. gracilis*. *PLOS ONE* **8**, e77011 (2013). doi: 10.1371/journal.pone.0077011; pmid: 24167556
30. Companion Web site: Figure W8, Figure W9, Figure W10, and Figure W14 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
31. F. Burki, P. J. Keeling, Rhizaria. *Curr. Biol.* **24**, R103–R107 (2014). doi: 10.1016/j.cub.2013.12.025; pmid: 24502779
32. N. R. Swanberg, thesis, Massachussetts Institute of Technology (1974).
33. I. Probert *et al.*, *Brandtodinium* gen. nov. and *B. nutricula* comb. Nov. (Dinophyceae), a dinoflagellate commonly found in symbiosis with polycystine radiolarians. *J. Phycol.* **50**, 388–399 (2014). doi: 10.1111/jpy.12174
34. S. von der Heyden, E. E. Chao, K. Vickerman, T. Cavalier-Smith, Ribosomal RNA phylogeny of bodonid and diplonemid flagellates and the evolution of euglenozoa. *J. Eukaryot. Microbiol.* **51**, 402–416 (2004). doi: 10.1111/j.1550-7408.2004.tb00387.x; pmid: 15352322
35. E. Schnepf, Light and electron microscopical observations in *Rhynchopus coscinodiscivorus* spec. nov., a Colorless, phagotrophic euglenozoon with concealed flagella. *Arch. Protistenkd.* **144**, 63–74 (1994). doi: 10.1016/S0003-9365(11)80225-3
36. M. R. Dennett, Video plankton recorder reveals high abundances of colonial Radiolaria in surface waters of the central North Pacific. *J. Plankton Res.* **24**, 797–805 (2002). doi: 10.1093/plankt/24.8.797
37. L. Stemmann *et al.*, Global zoogeography of fragile macrozooplankton in the upper 100-1000 m inferred from the underwater video profiler. *ICES J. Mar. Sci.* **65**, 433–442 (2008). doi: 10.1093/icesjms/fsn010
38. A. F. Michaels, D. A. Caron, N. R. Swanberg, F. A. Howse, C. M. Michaels, Planktonic sarcodines (Acantharia, Radiolaria, Foraminifera) in surface waters near Bermuda: Abundance, biomass and vertical flux. *J. Plankton Res.* **17**, 131–163 (1995). doi: 10.1093/plankt/17.1.131
39. E. Haeckel, "Report on the Radiolaria collected by H.M.S. Challenger during the years 1873-1876" in *Report on the Scientific Results of the Voyage of* H.M.S. Challenger *During the Years 1873-76. Zoology.* (Neill, Edinburgh, 1887).
40. R. Siano, M. Montresor, I. Probert, F. Not, C. de Vargas, *Pelagodinium* gen. nov. and *P. béii* comb. nov., a dinoflagellate symbiont of planktonic foraminifera. *Protist* **161**, 385–399 (2010). doi: 10.1016/j.protis.2010.01.002; pmid: 20149979
41. J. Decelle *et al.*, An original mode of symbiosis in open ocean plankton. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 18000–18005 (2012). doi: 10.1073/pnas.1212303109; pmid: 23071304
42. Y. Shaked, C. de Vargas, Pelagic photosymbiosis: rDNA assessment of diversity and evolution of dinoflagellate symbionts and planktonic foraminiferal hosts. *Mar. Ecol. Prog. Ser.* **325**, 59–71 (2006). doi: 10.3354/meps325059
43. J. Decelle, New perspectives on the functioning and evolution of photosymbiosis in plankton: Mutualism or parasitism? *Commun. Integr. Biol.* **6**, e24560 (2013). doi: 10.4161/cib.24560; pmid: 23986805
44. R. Siano *et al.*, Distribution and host diversity of Amoebophryidae parasites across oligotrophic waters of the Mediterranean Sea. *Biogeosciences* **8**, 267–278 (2011). doi: 10.5194/bg-8-267-2011
45. D. Coats, M. Park, Parasitism of photosynthetic dinoflagellates by three strains of *Amoebophyra* (Dinophyta): Parasite survival, infectivity, generation time, and host specificity. *J. Phycol.* **528**, 520–528 (2002). doi: 10.1046/j.1529-8817.2002.01200.x
46. K. E. Wommack, R. R. Colwell, Virioplankton: Viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **64**, 69–114 (2000). doi: 10.1128/MMBR.64.1.69-114.2000; pmid: 10704475
47. A. Skovgaard, Dirty tricks in the plankton: Diversity and role of marine parasitic protists. *Acta Protozool.* **53**, 51–62 (2014).
48. J. Bråte *et al.*, Radiolaria associated with large diversity of marine alveolates. *Protist* **163**, 767–777 (2012). doi: 10.1016/j.protis.2012.04.004; pmid: 22658831
49. T. R. Bachvaroff, S. Kim, L. Guillou, C. F. Delwiche, D. W. Coats, Molecular diversity of the syndinean genus *Euduboscquella* based on single-cell PCR analysis. *Appl. Environ. Microbiol.* **78**, 334–345 (2012). doi: 10.1128/AEM.06678-11; pmid: 22081578
50. S. Rueckert, T. G. Simdyanov, V. V. Aleoshin, B. S. Leander, Identification of a divergent environmental DNA sequence clade using the phylogeny of gregarine parasites (Apicomplexa) from crustacean hosts. *PLOS ONE* **6**, e18163 (2011). doi: 10.1371/journal.pone.0018163; pmid: 21483868
51. A. Skovgaard, S. A. Karpov, L. Guillou, The parasitic dinoflagellates *Blastodinium* spp. inhabiting the gut of marine, planktonic copepods: Morphology, ecology, and unrecognized species diversity. *Front. Microbiol.* **3**, 305 (2012). doi: 10.3389/fmicb.2012.00305; pmid: 22973263
52. H. McCallum *et al.*, Does terrestrial epidemiology apply to marine systems? *Trends Ecol. Evol.* **19**, 585–591 (2004). doi: 10.1016/j.tree.2004.08.009

53. Companion Web site: detailed Material and Methods, Database W9, and Figure W11 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
54. Companion Web site: Figure W12, Figure W13, Database W7, and Database W8 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
55. M. Holyoak, M. A. Leibold, R. D. Holt, Eds., *Metacommunities: Spatial Dynamics and Ecological Communities* (University of Chicago Press, Chicago, 2005).
56. L. G. M. Baas Becking, *Geobiologie of Inleiding tot de Milieukunde* (W. P. Van Stockum and Zoon, The Hague, Netherlands, 1934).
57. K. T. C. A. Peijnenburg, E. Goetze, High evolutionary potential of marine zooplankton. *Ecol. Evol.* **3**, 2765–2781 (2013). doi: 10.1002/ece3.644; pmid: 24567838
58. V. Smetacek, Microbial food webs. The ocean's veil. *Nature* **419**, 565 (2002). doi: 10.1038/419565a; pmid: 12374956
59. H. ter Steege *et al.*, Hyperdominance in the Amazonian tree flora. *Science* **342**, 1243092 (2013). doi: 10.1126/science.1243092; pmid: 24136971
60. D. Vaulot, K. Romari, F. Not, Are autotrophs less diverse than heterotrophs in marine picoplankton? **10**, 266–267 (2002).
61. A. H. Knoll, Paleobiological perspectives on early eukaryotic evolution. *Cold Spring Harb. Perspect. Biol.* **6**, 1–14 (2014). doi: 10.1101/cshperspect.a016121; pmid: 24384569
62. V. Smetacek, A watery arms race. *Nature* **411**, 745 (2001). doi: 10.1038/35081210; pmid: 11459035
63. S. Sunagawa *et al.*, Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
64. M. J. Oliver, D. Petrov, D. Ackerly, P. Falkowski, O. M. Schofield, The mode and tempo of genome size evolution in eukaryotes. *Genome Res.* **17**, 594–601 (2007). doi: 10.1101/gr.6096207; pmid: 17420184
65. H. Abida *et al.*, Bioprospecting marine plankton. *Mar. Drugs* **11**, 4594–4611 (2013). doi: 10.3390/md11114594; pmid: 24240981
66. G. Lima-Mendez *et al.*, Determinants of community structure in the global plankton interactome. *Science* **348**, 1262073 (2015).
67. K. D. Lafferty, A. P. Dobson, A. M. Kuris, Parasites dominate food web links. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 11211–11216 (2006). doi: 10.1073/pnas.0604755103; pmid: 16844774
68. L. A. Amaral-Zettler, E. A. McCliment, H. W. Ducklow, S. M. Huse, A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLOS ONE* **4**, e6372 (2009). doi: 10.1371/journal.pone.0006372; pmid: 19633714
69. L. Guillou *et al.*, The Protist Ribosomal Reference database (PR^2): A catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* **41**, D597–D604 (2013). doi: 10.1093/nar/gks1160; pmid: 23193267
70. R. Massana, J. del Campo, M. E. Sieracki, S. Audic, R. Logares, Exploring the uncultured microeukaryote majority in the oceans: Reevaluation of ribogroups within stramenopiles. *ISME J.* **8**, 854–866 (2014). doi: 10.1038/ismej.2013.204; pmid: 24196325
71. Companion Web site: Database W4 and Database W5 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
72. A. Godhe *et al.*, Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. *Appl. Environ. Microbiol.* **74**, 7174–7182 (2008). doi: 10.1128/AEM.01298-08; pmid: 18849462
73. F. Zhu, R. Massana, F. Not, D. Marie, D. Vaulot, Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol. Ecol.* **52**, 79–92 (2005). doi: 10.1016/j.femsec.2004.10.006; pmid: 16329895
74. Companion Web site: Text W2 and Figure W3 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
75. M. Kim, S. Nam, W. Shin, D. W. Coats, M. Park, *Dinophysis caudata* (dinophyceae) sequesters and retains plastids from the mixotrophic ciliate prey *Mesodinium rubrum*. *J. Phycol.* **48**, 569–579 (2012). doi: 10.1111/j.1529-8817.2012.01150.x

ACKNOWLEDGMENTS

We thank the following people and sponsors for their commitment: CNRS (in particular, the GDR3280); EMBL; Genoscope/CEA; UPMC; VIB; Stazione Zoologica Anton Dohrn; UNIMIB; Rega Institute; KU Leuven; Fund for Scientific Research – The French Ministry of Research, the French Government "Investissements d'Avenir" programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), and MEMO LIFE (ANR-10-LABX-54); PSL* Research University (ANR-11-IDEX-0001-02); ANR (projects POSEIDON/ANR-09-BLAN-0348, PROMETHEUS/ANR-09-PCS-GENM-217, PHYTBACK/ANR-2010-1709-01, and TARA-GIRUS/ANR-09-PCS-GENM-218); EU FP7 (MicroB3/No.287589, IHMS/HEALTH-F4-2010-261376); European Research Council Advanced Grant Awards to C. Bowler (Diatomite:294823); Gordon and Betty Moore Foundation grant 3790 to M.B.S.; Spanish Ministry of Science and Innovation grant CGL2011-26848/BOS MicroOcean PANGENOMICS and TANIT (CONES 2010-0036) grant from the Agency for Administration of University and Research Grants (AGAUR) to S.G.A.; and Japan Society for the Promotion of Science KAKENHI grant 26430184 to H.O. We also thank the following for their support and commitment: A. Bourgois, E. Bourgois, R. Troublé, Région Bretagne, G. Ricono, the Veolia Environment Foundation, Lorient Agglomération, World Courier, Illumina, the Electricité de France Foundation, Fondation pour la Recherche sur la Biodiversité, the Prince Albert II de Monaco Foundation, and the *Tara* schooner and its captains and crew. We thank MERCATOR-CORIOLIS and ACRI-ST for providing daily satellite data during the expedition. We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries who granted sampling permissions. *Tara* Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org). We also acknowledge assistance from European Bioinformatics Institute (EBI) (in particular, G. Cochrane and P. ten Hoopen) as well as the EMBL Advanced Light Microscopy Facility (in particular, R. Pepperkok). We thank F. Gaill, B. Kloareg, F. Lallier, D. Boltovskoy, A. Knoll, D. Richter, and E. Médard for help and advice on the manuscript. We declare that all data reported herein are fully and freely available from the date of publication, with no restrictions, and that all of the samples, analyses, publications, and ownership of data are free from legal entanglement or restriction of any sort by the various nations whose waters the *Tara* Oceans expedition sampled in. Data described herein are available at http://taraoceans.sb-roscoff.fr/EukDiv/, at EBI under the project IDs PRJEB402 and PRJEB6610, and at PANGAEA (see table S1). The data release policy regarding future public release of *Tara* Oceans data is described in (*12*). All authors approved the final manuscript. This article is contribution number 24 of *Tara* Oceans. The supplementary materials contain additional data.

Tara Oceans Coordinators

Silvia G. Acinas,[1] Peer Bork,[2] Emmanuel Boss,[3] Chris Bowler,[4] Colomban de Vargas,[5,6] Michael Follows,[7] Gabriel Gorsky,[8,9] Nigel Grimsley,[10,11] Pascal Hingamp,[12] Daniele Iudicone,[13] Olivier Jaillon,[14,15,16] Stefanie Kandels-Lewis,[2,17] Lee Karp-Boss,[3] Eric Karsenti,[4,17] Uros Krzic,[18] Fabrice Not,[5,6] Hiroyuki Ogata,[19] Stephane Pesant,[20,21] Jeroen Raes,[22,23,24] Emmanuel G. Reynaud,[25] Christian Sardet,[26,27] Mike Sieracki,[28] Sabrina Speich,[29,30] Lars Stemmann,[8] Matthew B. Sullivan,[31]* Shinichi Sunagawa,[2] Didier Velayoudon,[32] Jean Weissenbach,[14,15,16] Patrick Wincker[14,15,16]

[1]Department of Marine Biology and Oceanography, ICM-CSIC, Passeig Marítim de la Barceloneta, 37-49, Barcelona E08003, Spain. [2]Structural and Computational Biology, EMBL, Meyerhofstraße 1, 69117 Heidelberg, Germany. [3]School of Marine Sciences, University of Maine, Orono, ME 04469, USA. [4]Ecole Normale Supérieure, IBENS, and Inserm U1024, and CNRS UMR 8197, Paris, F-75005 France. [5]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [6]Sorbonne Universités, UPMC Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [7]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [8]CNRS, UMR 7093, LOV, Observatoire Océanologique, F-06230, Villefranche-sur-Mer, France. [9]Sorbonne Universités, UPMC Paris 06, UMR 7093, LOV, Observatoire Océanologique, F-06230, Villefranche-sur-Mer, France. [10]CNRS UMR 7232, BIOM, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [11]Sorbonne Universités Paris 06, OOB UPMC, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [12]Aix Marseille Université, CNRS IGS, UMR 7256, 13288 Marseille, France. [13]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [14]CEA, Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France. [15]CNRS, UMR 8030, CP5706 Evry, France. [16]Université d'Evry, UMR 8030, CP5706 Evry, France. [17]Directors' Research, EMBL, Meyerhofstraße 1, 69117 Heidelberg, Germany. [18]Cell Biology and Biophysics, EMBL, Meyerhofstraße 1, 69117 Heidelberg, Germany. [19]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan. [20]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany. [21]MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany. [22]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [23]Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. [24]Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [25]Earth Institute, University College Dublin, Dublin, Ireland. [26]CNRS, UMR 7009 Biodev, Observatoire Océanologique, F-06230 Villefranche-sur-Mer, France. [27]Sorbonne Universités, UPMC Univ Paris 06, UMR 7009 Biodev, F-06230 Observatoire Océanologique, Villefranche-sur-Mer, France. [28]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME 04544, USA. [29]Department of Geosciences, LMD, Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris, Cedex 05, France. [30]Laboratoire de Physique des Océan UBO-IUEM Place Copernic 29820 Plouzané, France. [31]Department of Ecology and Evolutionary Biology, University of Arizona, 1007 East Lowell Street, Tucson, AZ 85721, USA. [32]DVIP Consulting, Sèvres, France.
*Present address: Department of Microbiology, Ohio State University, Columbus, OH 43210, USA.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/1261605/suppl/DC1
Table S1
Appendix S1

23 September 2014; accepted 27 February 2015
10.1126/science.1261605

D. Co-authored manuscript 2: Lima-Mendez et al, 2015

Lima-Mendez et al., 2015 :

I was in charge of providing all the interpretation regarding diatom interactions in the network, including the *in silico* search for symbiotic interactions or parasitism, and their particular role in the PFT graph (**Figure 3**). I was also in charge of providing a table of comparison for diatom interactions between litterature, co-occurrence inference, and light microscopy evidence provided by Sébastien Colin as a means of method assessment, as well as contributing to the literature database used in the Interactome paper. Finally, even though unsuccessful, the search for diatom-related interactions in the fixed samples was done, based on selected inferred interactions.

OCEAN PLANKTON

Determinants of community structure in the global plankton interactome

Gipsi Lima-Mendez,[1,2,3]* Karoline Faust,[1,2,3]* Nicolas Henry,[4,5]* Johan Decelle,[4,5] Sébastien Colin,[4,5,6] Fabrizio Carcillo,[1,2,3,7] Samuel Chaffron,[1,2,3] J. Cesar Ignacio-Espinosa,[8]† Simon Roux,[8]† Flora Vincent,[2,6] Lucie Bittner,[4,5,6,9] Youssef Darzi,[2,3] Jun Wang,[1,2] Stéphane Audic,[4,5] Léo Berline,[10,11] Gianluca Bontempi,[7] Ana M. Cabello,[12] Laurent Coppola,[10,11] Francisco M. Cornejo-Castillo,[12] Francesco d'Ovidio,[13] Luc De Meester,[14] Isabel Ferrera,[12] Marie-José Garet-Delmas,[4,5] Lionel Guidi,[10,11] Elena Lara,[12] Stéphane Pesant,[15,16] Marta Royo-Llonch,[12] Guillem Salazar,[12] Pablo Sánchez,[12] Marta Sebastian,[12] Caroline Souffreau,[14] Céline Dimier,[4,5,6] Marc Picheral,[10,11] Sarah Searson,[10,11] Stefanie Kandels-Lewis,[17,18] Tara Oceans coordinators‡ Gabriel Gorsky,[10,11] Fabrice Not,[4,5] Hiroyuki Ogata,[19] Sabrina Speich,[20,21] Lars Stemmann,[10,11] Jean Weissenbach,[22,23,24] Patrick Wincker,[22,23,24] Silvia G. Acinas,[12] Shinichi Sunagawa,[17] Peer Bork,[17,25] Matthew B. Sullivan,[8]† Eric Karsenti,[6,18]§ Chris Bowler,[6]§ Colomban de Vargas,[4,5]§ Jeroen Raes[1,2,3]§

Species interaction networks are shaped by abiotic and biotic factors. Here, as part of the *Tara* Oceans project, we studied the photic zone interactome using environmental factors and organismal abundance profiles and found that environmental factors are incomplete predictors of community structure. We found associations across plankton functional types and phylogenetic groups to be nonrandomly distributed on the network and driven by both local and global patterns. We identified interactions among grazers, primary producers, viruses, and (mainly parasitic) symbionts and validated network-generated hypotheses using microscopy to confirm symbiotic relationships. We have thus provided a resource to support further research on ocean food webs and integrating biological components into ocean models.

The structure of oceanic ecosystems results from the complex interplay between resident organisms and their environment. In the world's largest ecosystem, oceanic plankton (composed of viruses, prokaryotes, microbial eukaryotes, phytoplankton, and zooplankton) form trophic and symbiotic interaction networks (*1*–*4*) that are influenced by environmental conditions. Ecosystem structure and composition are governed by abiotic as well as biotic factors. The former include environmental conditions and nutrient availability (*5*), whereas the latter include grazing, pathogenicity, and parasitism (*6*, *7*). Historically, abiotic factors have been considered to have a stronger effect, but recently, appreciation for biotic factors is growing (*8*, *9*). We sought to develop a quantitative understanding of biotic and abiotic interactions in natural systems in which the organisms are taxonomically and trophically diverse (*10*). We used sequencing technologies to profile communities across trophic levels, organismal sizes, and geographic ranges and to predict organismal interactions across biomes based on co-occurrence patterns (*11*). Previous efforts addressing these issues have provided insights on the structure (*12*, *13*) and dynamics of microbial communities (*14*–*16*).

We analyzed data from 313 plankton samples the *Tara* Oceans expedition (*17*) derived from seven size-fractions covering collectively 68 stations at two depths across eight oceanic provinces (table S1). The plankton samples spanned sizes that include organisms from viruses to small metazoans. We derived viral, prokaryotic, and eukaryotic abundance profiles from clusters of metagenomic contigs, Illumina-sequenced metagenomes ($_{mi}$tags), and 18*S* ribosomal DNA (rDNA) V9 sequences, respectively (table S1) (*10*, *18*, *19*) and collected environmental data from on-site and satellite measurements (*17*, *20*, *21*). We used network inference methods and machine-learning techniques so as to disentangle biotic and abiotic signals shaping ocean plankton communities and to construct an interactome that described the network of interactions among photic zone plankton groups. We used the interactome to focus on specific relationships, which we validated through microscopic analysis of symbiont pairs and in silico analysis of phage-host pairings.

Evaluating the effect of abiotic and biotic factors on community structure

We first reassessed the effects of environment and geography on community structure. Using variation partitioning (*22*), we found that on average, the percentage of variation in community composition explained by environment alone was 18%, by environment combined with geography 13%, and by geography alone only 3% (*23*, *24*). In addition, we built random forest-based models (*25*) in order to predict abundance profiles of the Operational Taxonomic Units (OTU) using (i) OTUs alone, (ii) environmental variables alone, and (iii) OTUs and environmental variables combined and tested for each OTU whether one of the three approaches outcompeted the other. These analyses revealed that 95% of the OTU-only models are more accurate in predicting OTU abundances than environmental variable models, and that combined models were no better than the OTU-only models (*26*, *27*). This suggests that abiotic factors have a more limited effect on community structure than previously assumed (*8*).

To study the role of biotic interactions, we developed a method with which to identify robust species associations in the context of environmental conditions. Twenty-three taxon-taxon and taxon-environment co-occurrence networks were constructed based on 9292 taxa, representing the combinations of two depths, seven organismal

[1]Department of Microbiology and Immunology, Rega Institute KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [2]VIB Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. [3]Department of Applied Biological Sciences (DBIT), Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [4]Station Biologique de Roscoff, CNRS, UMR 7144, Place Georges Teissier, 29680 Roscoff, France. [5]Sorbonne Universités, Université Pierre et Marie Curie (UPMC) Université Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [6]Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), Inserm U1024, CNRS UMR 8197, Paris, F-75005 France. [7]Interuniversity Institute of Bioinformatics in Brussels (IB)^2, ULB Machine Learning Group, Computer Science Department, Université Libre de Bruxelles (ULB), Brussels, Belgium. [8]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, 85721, USA. [9]Institut de Biologie Paris-Seine, CNRS FR3631, F-75005, Paris, France. [10]CNRS, UMR 7093, Laboratoire d'Océanographie de Villefranche (LOV), Observatoire Océanologique, F-06230 Villefranche-sur-mer, France. [11]Sorbonne Universités, UPMC Paris 06, UMR 7093, Laboratoire d'Océanographie de Villefranche (LOV), Observatoire Océanologique, F-06230 Villefranche-sur-mer, France. [12]Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM)–Consejo Superior de Investigaciones Científicas (CSIC), Pg. Marítim de la Barceloneta, 37-49, Barcelona E08003, Spain. [13]Sorbonne Universités, UPMC, Université Paris 06, CNRS–Institut pour la Recherche et le Développement–Muséum National d'Histoire Naturelle, Laboratoire d'Océanographie et du Climat: Expérimentations et Approches Numériques (LOCEAN) Laboratory, 4 Place Jussieu, 75005, Paris, France. [14]KU Leuven, Laboratory of Aquatic Ecology, Evolution and Conservation, Charles Deberiotstraat 32, 3000 Leuven. [15]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Hochschulring 18, 28359 Bremen, Germany. [16]MARUM, Center for Marine Environmental Sciences, University of Bremen, Hochschulring 18, 28359 Bremen, Germany. [17]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [18]Directors' Research, European Molecular Biology Laboratory, Heidelberg, Germany. [19]Institute for Chemical Research, Kyoto University, Gokasho, Uji, 611-0011 Kyoto, Japan. [20]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France. [21]Laboratoire de Physique des Océan, Université de Bretagne Occidentale (UBO)–Institut Universaire Européen de la Mer (IUEM), Palce Copernic, 29820 Polouzané, France. [22]Commissariat à l'Énergie Atomique (CEA), Genoscope, 2 rue Gaston Crémieux, 91000 Evry, France. [23]CNRS, UMR 8030, 2 rue Gaston Crémieux, 91000 Evry, France. [24]Université d'Evry, UMR 8030, CP5706 Evry, France. [25]Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany.

*These authors contributed equally to this work. †Present address: Department of Microbiology, Ohio State University, Columbus, OH 43210, USA. ‡*Tara* Oceans coordinators and affiliations are listed at the end of this manuscript. §Corresponding author. E-mail: jeroen.raes@vib-kuleuven.be (J.R.); vargas@sb-roscoff.fr (C.d.V.); cbowler@biologie.ens.fr (C.B.); karsenti@embl.de (E.K.)

size ranges, and four organismal domains (Bacteria, Archaea, Eukarya, and viruses) (*28*). To reduce noise and thus false-positive predictions, we restricted our analysis to taxa present in at least 20% of the samples and used conservative statistical cutoffs. We merged the individual networks into a global network, which features a total of 127,995 distinct edges, of which 92,633 are taxon-taxon edges and 35,362 are taxon-environment edges (Table 1). Node degree does not depend on the abundance of the node (*28*). As such, this network represents a resource with which to examine species associations in the global oceans (*28*–*31*).

Next, we assessed how many of the taxon links represented "niche effects" driven by geography or environment (such as when taxa respond similarly to a common environmental condition). We examined motifs consisting of two correlated taxa that also correlate with at least one common environmental parameter ("environmental triplets" to identify associations that were driven by environment) using three approaches [interaction information, sign pattern analysis, and network deconvolution (*32*)]. We identified 29,912 taxon-taxon-environment associations (32.3% of total). Among environmental factors, we found that PO_4, temperature, NO_2, and mixed-layer depth were frequent drivers of network connections (Fig. 1A). Although the three methodologies pinpoint indirect associations, only interaction information directly identifies synergistic effects in these biotic-abiotic triplets. Exploiting this property, we disentangled the 29,912 environment-affected associations into 11,043 edges driven solely by abiotic factors (excluded from the network for the remainder of the study) (*31*, *33*) and 18,869 edges whose dependencies result from biotic-abiotic synergistic effects. Thus, we find that a minority of associations can be explained by an environmental factor.

Evaluation of predicted interactions

Co-occurrence techniques have heretofore mainly been applied to bacteria. We detected eukaryotic interactions on the basis of analysis of sequences at the V9 hypervariable region of the 18*S* ribosomal RNA (rRNA) gene. We built a literature-curated collection (*34*) of 574 known symbiotic interactions (including both parasitism and mutualism) in marine eukaryotic plankton (*30*, *35*). From 43 genus-level interactions represented by OTUs in the abundance preprocessed input matrices, we found 42% (18 genus pairs; 47% when limiting to parasitic interactions) represented in our reference list. The probability of having found each of these interactions by chance alone was <0.01 (Fisher exact test, average $P = 4^{-3}$, median $P = 5e^{-7}$). On the basis of this sensitivity and a false discovery rate averaging to 9% (computed from null models), we estimate the number of interactions among eukaryotes present in our filtered input matrices to be between 53,000 and 139,000. Most of the false-negative interactions were due to the strict filtering rules we used to avoid false positives; this hampers detection when, for example, interactions are facultative or when interaction partners may vary among closely related groups depending on oceanic region (*4*). False positives could represent indirect interactions between species (bystander effects) or environmental effects caused by factors not captured in this study (*36*, *37*).

Biotic interactions within and across kingdoms

The integrated network contained 81,590 predicted biotic interactions (*30*) that were nonrandomly distributed within and between size fractions (Fig. 1, B and C) (*38*). Positive associations outnumbered mutual exclusions (72% versus 28%), and we observed a nonrandom edge distribution with regard to phylogeny (Fig. 2A), with most associations derived from syndiniales and other dinoflagellates (examples are shown in

Table 1. Properties of the merged taxon network. The positive subset of the network was clustered with the leading eigen vector algorithm (*91*).

Nodes	Edges	Positive edges (%)	Negative edges	Average clustering coefficient	Average path length	Diameter	Average betweenness	Modularity of positive network	Number of modules in positive network
9169	92,633	68,856 (74.33)	23,777	0.229	3.43	12	11024	0.51	51

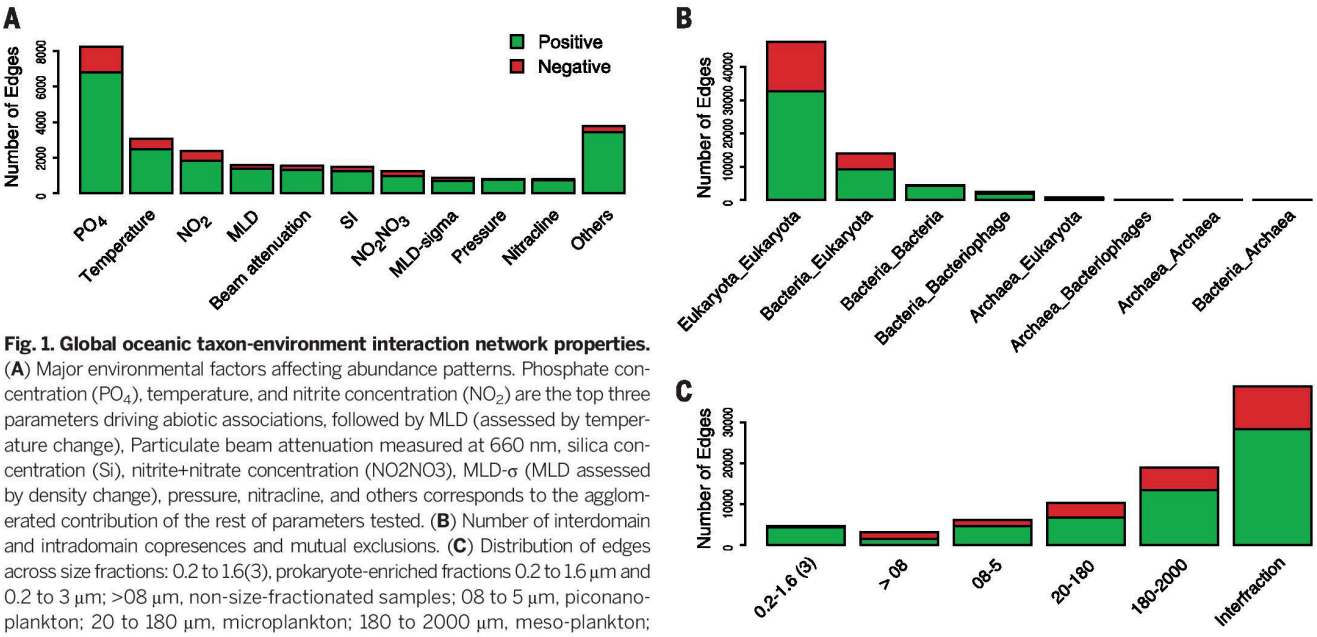


Fig. 1. Global oceanic taxon-environment interaction network properties. (**A**) Major environmental factors affecting abundance patterns. Phosphate concentration (PO_4), temperature, and nitrite concentration (NO_2) are the top three parameters driving abiotic associations, followed by MLD (assessed by temperature change), Particulate beam attenuation measured at 660 nm, silica concentration (Si), nitrite+nitrate concentration (NO2NO3), MLD-σ (MLD assessed by density change), pressure, nitracline, and others corresponds to the agglomerated contribution of the rest of parameters tested. (**B**) Number of interdomain and intradomain copresences and mutual exclusions. (**C**) Distribution of edges across size fractions: 0.2 to 1.6(3), prokaryote-enriched fractions 0.2 to 1.6 μm and 0.2 to 3 μm; >08 μm, non-size-fractionated samples; 08 to 5 μm, piconanoplankton; 20 to 180 μm, microplankton; 180 to 2000 μm, meso-plankton; interfrac, includes interfraction networks 08 to 5 μm versus 20 to 180 μm, 08 to 5 μm versus 180 to 2000 μm, 20 to 180 μm versus 180 to 2000 μm, and 0.2 to 1.6(3) μm versus ≤ 0.2 μm (virus-enriched fraction).

Fig. 3A), and exclusions involving arthropods. Certain combinations of phylogenetic groups are overrepresented (*39*). For instance, we found a clade of syndiniales [the MALV-II Clade 1 belonging to *Amoebophrya* (*3*)] enriched in positive associations with tintinnids ($P = 2^{-4}$), which are among the most abundant ciliates in marine plankton (*40*). The tintinnid *Xystonella lohmani* was described in 1964 to be infected by *Amoebophrya tintinnis* (*41*), and tintinnids can feed on *Amoebophrya* free-living stages (*42*). Other found host-parasite associations included the copepod parasites *Blastodinium*, *Ellobiopsis*, and *Vampyrophrya* (*41*, *43*–*45*).

On the other hand, *Maxillopoda*, *Bacillariophyceae*, and collodarians, three groups of relatively large sized organisms whose biomass can dominate planktonic ecosystems, are rich in negative associations among them (*33*). Collodarians and copepods are abundant in, respectively, the oligotrophic tropical and eutrophic and mesotrophic temperate systems (*10*, *46*). The decoupling of phyto- and zooplankton in open oceans by diatoms anticorrelating to copepods (*47*, *48*) is attributed to growth rate differences and to the diatom production of compounds harmful to their grazers (*49*). The combination of these

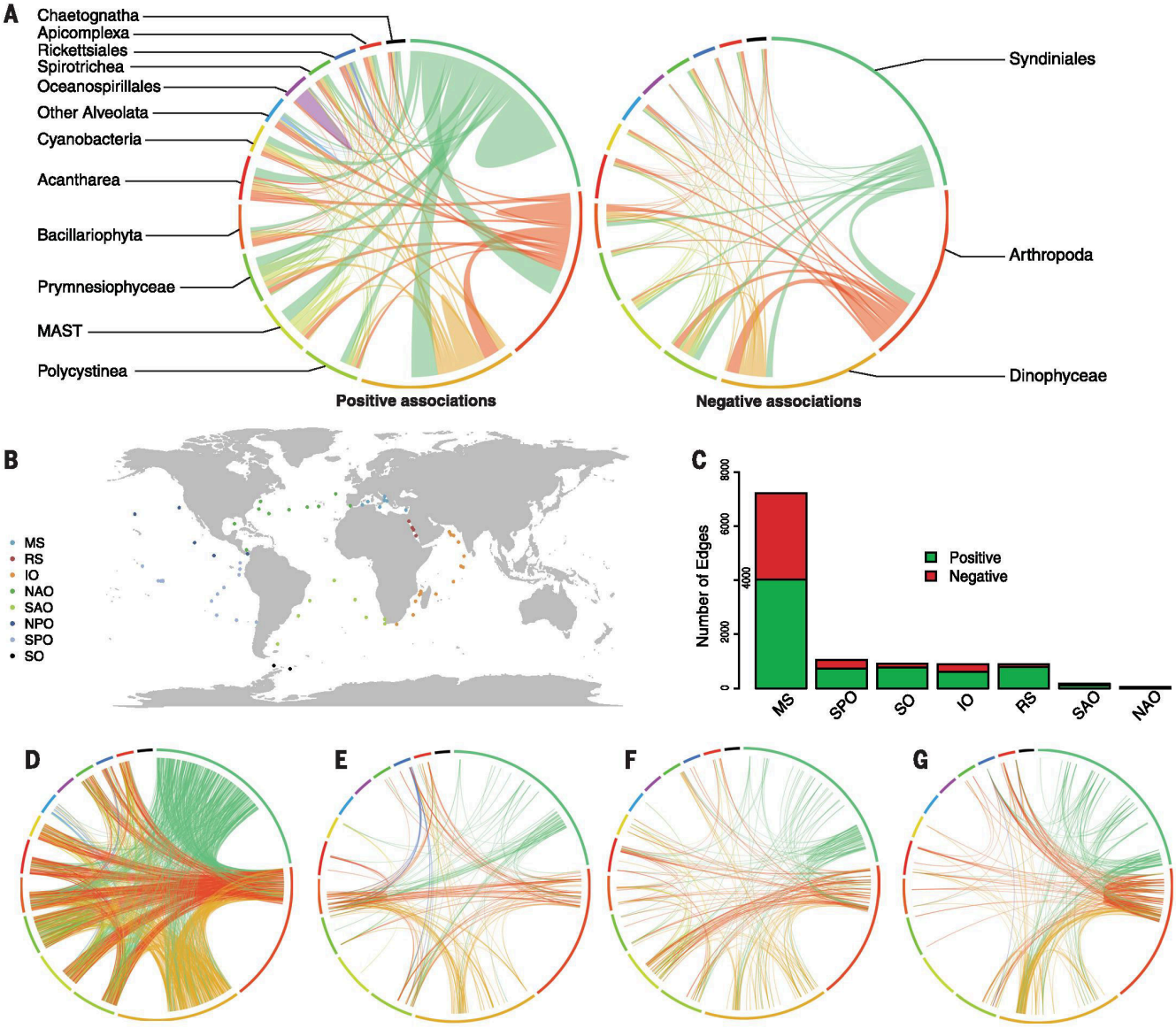


Fig. 2. Taxonomic and geographic patterns within the co-occurrence network. (**A**) Top 15 interacting taxon groups depicted as colored segments in a CIRCOS plot, in which ribbons connecting two segments indicate co-presence and exclusion links, on the left and right, respectively. Size of the ribbon is proportional to the number of links (copresences and exclusions) between the OTUs assigned to the respective segments, and color is segment (of the two involved) with the more total links. Links are dominated by the obligate parasites syndiniales and by Arthropoda and Dinophyceae. (**B**) *Tara* Oceans sampling stations grouped by oceanic provinces. (**C**) Frequency of local co-occurrence patterns across the oceanic provinces, showing that most local patterns are located in MS. (**D** to **G**) Taxonomic patterns of co-occurrences across MS (D), SPO (E), IO (F), and RS (G). Edges are represented as ribbons between barcodes grouped into their taxonomic order as in (A). Links sharing the same segment are affiliated to the same taxon (Order), showing that the connectivity patterns across taxa are conserved at high taxonomic ranks. The local specificity of interactions at higher resolution (OTUs) is apparent by thin ribbons (edge resolution), with different starts, and end positions (different OTUs) within the shared (taxon) segment, section color, and ordering correspond to those in (A). SO-specific associations are mainly driven by bacterial interactions (*53*).

effects could lie at the basis of this observation, which contrasts with other free-living autotrophs represented in the network (cyanobacteria and prymnesiophytes), which display primarily positive associations (Fig. 2A).

Cross-kingdom associations between Bacteria and Archaea were limited to 24 mutual exclusions. Within Archaea, Thermoplasmatales (Marine Group II) co-occur with several phytoplankton clades. Links between Bacteria and protists recovered five out of eight recently discovered interactions from protist single-cell sequencing (*50*). Associations between Diatoms and Flavobacteria agreed with their described symbioses (*51*). We also observed co-ocurrence of uncultured dinoflagellates with members of Rhodobacterales (*Ruegeria*), which is in agreement with a symbiosis between *Ruegeria* sp. TM1040 and *Pfiesteria piscicida* around the ability of *Ruegeria* to metabolize dinoflaggelate-produced dimethylsulfoniopropionate (*52*).

Global versus local associations

We further investigated whether our network was driven by global trends or is defined by local signals. To this aim, we divided our set of samples into seven main regions—Mediterranean

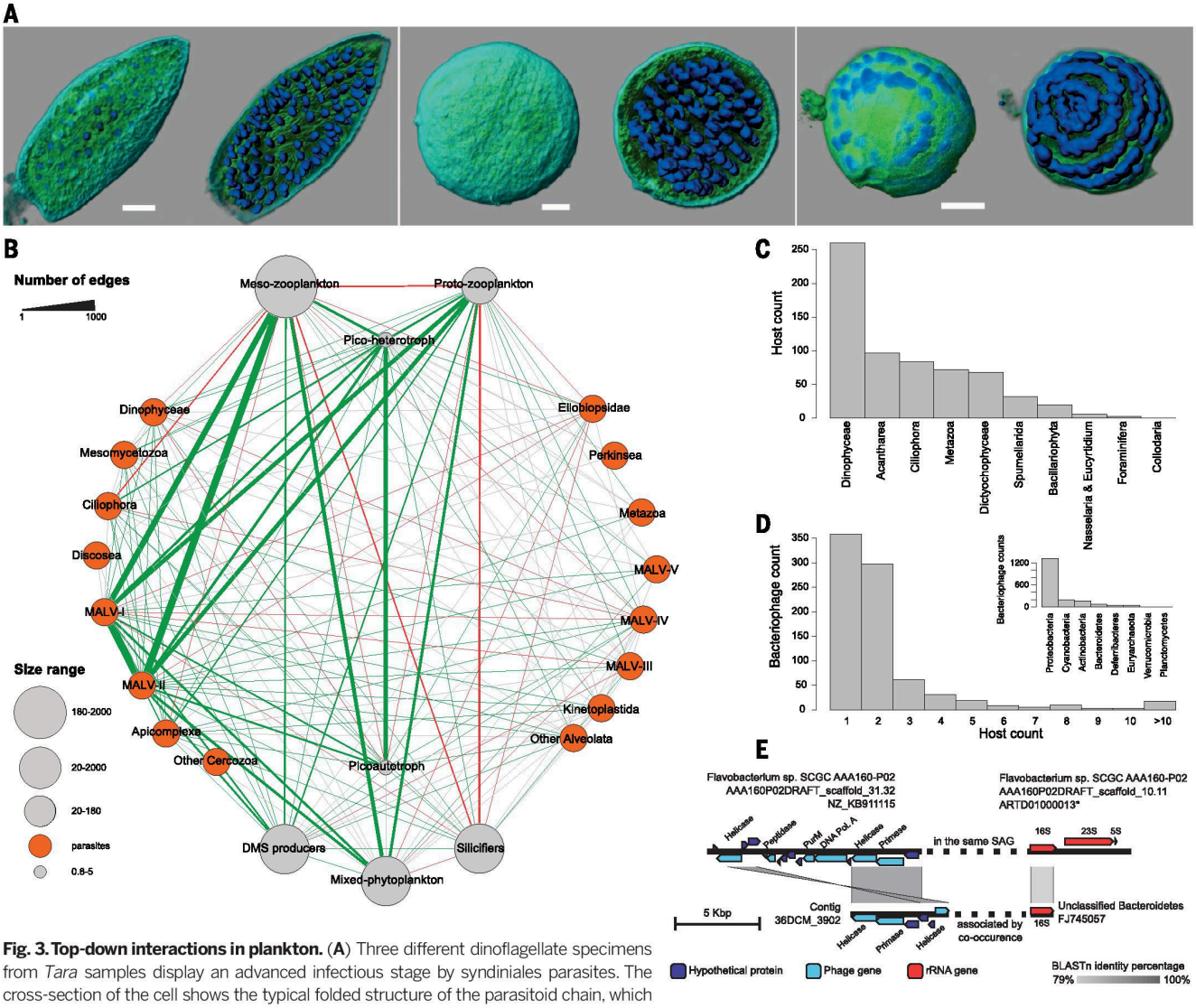


Fig. 3. Top-down interactions in plankton. (**A**) Three different dinoflagellate specimens from *Tara* samples display an advanced infectious stage by syndiniales parasites. The cross-section of the cell shows the typical folded structure of the parasitoid chain, which fills the entire host cell. Each nucleus (blue) of the coiled ribbon corresponds to a future free-living parasite. DNA is stained with Hoechst (dark blue), membranes are stained with DiOC6 (green), and specimen surface is light blue. Scale bar, 5 μm. (**B**) Subnetwork of metanodes that encapsulate barcodes affiliated to parasites or PFTs. The PFTs mapped onto the network are: phytoplankton DMS producers, mixed phytoplankton, phytoplankton silicifiers, pico-eukaryotic heterotrophs, proto-zooplankton and meso-zooplankton. Edge width reflects the number of edges in the taxon graph between the corresponding metanodes. Over-represented links (multiple-test corrected $P < 0.05$) are colored in green if they represent copresences and in red if they represent exclusions; gray means non-overrepresented combinations. When both copresences and exclusions were significant, the edge is shown as copresence. (**C**) Parasite connections within micro- and zooplankton groups. (**D**) Number of hosts per phage. (Inset) Phage associations to bacterial (target) phyla. (**E**) Putative Bacteroidetes viruses detected with co-occurence and detection in a single-cell genome (SAG). On the left are viral sequences from a Flavobacterium SAG (top) and *Tara* Oceans virome (bottom), displaying an average of 89% nucleotide identity. On the right is the correspondence between the ribosomal genes detected in the same SAG (top) and the 16S sequence associated to the *Tara* Oceans contig based on co-occurence (79% nucleotide identity). For clarity, a subset of contig ARTD0100013 only (from 10,000 to 16,000 nucleotides) is displayed. This sequence was also reverse-complemented. PurM, phosphoribosylaminoimidazole synthetase; DNA Pol. A, DNA polymerase A.

Sea (MS), Red Sea (RS), Indian Ocean (IO), South Atlantic (SAO), Southern Ocean (SO), South Pacific Ocean (SPO) and North Atlantic Ocean (NAO)—and assessed the "locality" of associations by comparing the score with or without that region. We found that association patterns were mostly driven by global trends because only 14% of edges were identified as local (Fig. 2, B and C). Approximately two thirds of local associations occur in MS (7215), followed by SPO (1058), whereas the rest are contributed by SO (901), IO (894), RS (889), SAO (163), and NAO (60) (Fig. 2, C to G). MS was the region with most sampling sites, which allowed us to recover more local patterns. Nevertheless, Fig. 2, C to G, shows that although the same major groups (order level) interact in both the global and local networks, each local site has its own specific interaction profile ($P < 1^{-8}$) (*33*, *39*, *53*).

Parasite impact on plankton functional types

Parasitic interactions are the most abundant pattern present in the network, which is also eminent by repeated microscopic observation of parasitic interactions from the *Tara* samples (Fig. 3A). We focused on predicted parasitic interactions and assessed their potential impact on biogeochemical processes by exploring a functional subnetwork (21,572 edges) of known and previously unidentified plankton parasites (*10*) together with classical "plankton functional types" (PFTs) (*54*). PFTs group taxa by trophic strategy (for example, autotrophs versus heterotrophs) and role in ocean biogeochemistry (Fig. 3A) (*55*). The relationship between the different PFTs (network density of 0.65) highlights strong dependencies between phytoplankton and grazers. We found that all PFTs are associated with parasites, but not always to the same extent. Most links involve syndiniales MALV-I and MALV-II clades associated to zooplankton and, to a lesser extent, to microphytoplankton (excluding diatoms). This emphasizes the role of alveolate parasitoids as top-down effectors of zooplankton and microphytoplankton population structure and functioning (*3*), although the latter group is also affected by grazing (*1*). The meso-planktonic networks contain known syndiniales targets (Dinophyceae, Ciliophora, Acantharia, and Metazoa) (Fig. 3B) (*56*). In large size fractions, we found interactions between known parasites and groups of organisms that in theory are too small to be their hosts (*57*); 32% of these associations involved the abundant and diverse marine stramenopiles (MASTs) and diplonemids (other Discoba and Diplonema) (*10*). Ecophysiology studies (*58*, *59*) suggest a parasitic role for these lineages. The association of these groups with other parasites would be explained by putative co-infection of the same hosts. Contrasting with the above observations, we found phytoplankton silicifiers (diatoms) displaying a variety of mutual exclusions. One possible interpretation of this is that diatom silicate exoskeletons (*60*) and toxic compound production (*49*) could act as efficient barriers against top-down pressures (*61*).

Phage-microbe associations

We investigated phage-microbe interactions, another major top-down process affecting global bacterial/archaeal community structure (*7*). Here, surface (SRF) and deep chlorophyll maximum (DCM) virus-bacteria networks revealed 1869 positive associations between viral populations and 7 of the 54 known bacterial phyla (specifically, Proteobacteria, Cyanobacteria, Actinobacteria, Bacteroidetes, Deferribacteres, Verrucomicrobia, and Planctomycetes), and one archaeal phylum (Euryarchaeota). These eight phyla represent most of abundant bacterial/archaeal groups across 37 investigated samples (Fig. 3D), suggesting that the networks are detecting abundant virus-host interactions. Additionally, these interactions include phyla of microbes lacking viral genomes in RefSeq databases including Verrucomicrobia, and nonextremophile Euryarchaeota, hinting at genomic sequences for understudied viral taxa (Fig. 3E) (*39*, *62*, *63*). Among the phage populations in the network, we found eight corresponding to phage sequences available in GenBank (>50% of genes with a >50% amino acid identity match). In all eight cases, the predicted host from the network corresponded to the annotated host family in the GenBank record, which is significantly higher than expected by chance ($P = 0.001$) (*62*).

Next, we evaluated viral host range, which is fundamental for predictive modeling and thus far largely limited to observations of cultured virus-host systems that insufficiently map complex community interactions (*64*). Our virus-host interaction data suggest that viruses are very host-specific: ~43% of the phage populations interact with only a single host OTU, and the remaining 57% interact with only a few, often closely related OTUs (Fig. 3D). These networks are modular at large scales (*65*), suggesting that viruses are host range–limited across large sections of host space. Nestedness analysis showed inconsistent results across algorithms.

Microscopic validation of predicted interactions

Our data predicted a photosymbiotic interaction between an acoel flatworm (*Symsagittifera* sp.) and a green microalga (*Tetraselmis* sp.). We validated this by means of laser scanning confocal microscopy (LSCM), three-dimensional (3D) reconstruction, and reverse molecular identification on flatworm specimens isolated from *Tara* Oceans preserved morphological samples. We observed microalgal cells (5 to 10 μm in diameter) within each of the 15 isolated acoel specimens (Fig. 4) (*66*). The 18*S* sequence from several sorted holobionts matched the metabarcode pair identified in the co-occurrence global network. Thus, molecular ecology, bioinformatics, and microscopic analysis can enable the discovery of marine symbioses.

Conclusions

The global ocean interactome can be used to predict the dynamics and structure of ocean ecosystems. The interactome reported here spans all three organismal domains and viruses. The analyses presented emphasize the role of top-down biotic interactions in the epipelagic zone. This data will inform future research to understand how symbionts, pathogens, predators, and parasites interact with their target organisms and will ultimately help elucidate the structure of the global food webs that drive nutrient and energy flow in the ocean.

Methods

Sampling

The sampling strategy used in the *Tara* Oceans expedition is described in (*67*), and samples used in the present study are listed in table S1 and http://doi.pangaea.de/10.1594/PANGAEA.840721. The *Tara* Oceans nucleotide sequences are available at the European Nucleotide Archive (ENA) under projects PRJEB402 and PRJEB6610.

Physical and environmental measurements

Physical and environmental measurements were carried out with a vertical profile sampling system (CTD-rosette) and data collected from Niskin bottles. We measured temperature, salinity, chlorophyll, CDOM fluorescence (fluorescence of the colored dissolved organic matter), particles abundance, nitrate concentration, and particle size distribution (using an underwater vision profiler). In addition, mean mixed-layer depth (MLD), maximum fluorescence, vertical maximum of the Brünt-Väisälä Frequency N ($s - 1$), vertical range of dissolved oxygen, and depth of nitracline were determined. Satellite altimetry provided the Okubo-Weiss parameter, Lyapunov exponent, mesoscale eddie retention, and sea-surface temperature (SST) gradients at eddie fronts (*19*). Data are available at http://www.pangaea.de (http://doi.pangaea.de/10.1594/PANGAEA.840718).

Abundance table construction

Prokaryotic 16*S* rDNA metagenomic reads were identified, annotated, and quantified from $_{mi}$tags) as described in (*68*) by using the SILVA v.115 database (*19*, *69*, *70*). The abundance table was normalized by using the summed read count per sample (*19*, *71*). Quality-checked V9 rDNA metabarcodes were clustered into swarms as in (*10*, *72*) and annotated by using the V9 PR2 database (*73*). PR2 barcodes were associated to fundamental trophic modes (auto- or heterotrophy) and symbiotic interactions (parasitism and mutualism) according to literature (Taxonomic and trophic mode annotations are available at http://doi.pangaea.de/10.1594/PANGAEA.843018 and http://doi.pangaea.de/10.1594/PANGAEA.843022). Swarm abundance and normalization was performed as in (*10*, *72*). Bacteriophage metagenomes were obtained from the < 0.2-μm fractions for 48 samples, and contigs were annotated and quantified as in (*18*). The abundance matrix was normalized by means of total sample read count and contig length.

In all cases, only OTUs with relative abundance $> 1^{-8}$ and detected in at least 20% of samples were retained. Because sample number in the input tables ranged from 17 to 63, prevalence thresholds varied (from 22 to 40%). The sum of all filtered OTU relative abundances was kept in the tables to preserve proportions. Abundance tables

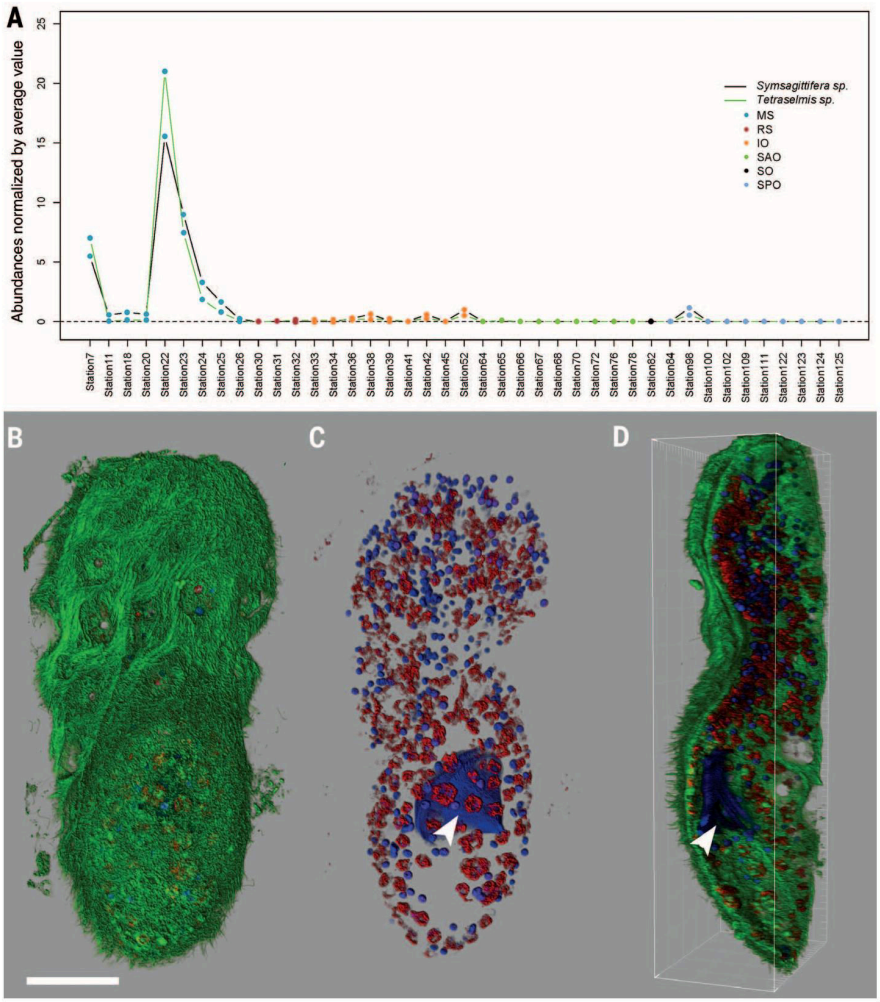


Fig. 4. Experimental validation of network-predicted interaction (photosymbiosis). Guided by the predictions from the co-occurrence network and abundance patterns, acoel flatworms (*Symsagittifera* sp.) together with their photosynthetic green microalgal endosymbionts (*Tetraselmis* sp.) were collected in microplankton samples from *Tara* Oceans Station 22 in the Mediterranean Sea. Pictures show a 3D reconstructed specimen from LSCM images [green channel, cellular membranes (DiOC6); blue channel, DNA and the nuclei (Hoechst33342); red channel: chlorophyll autofluorescence]. (**A**) Co-occurrence plot of *Symsagittifera*- and *Tetraselmis*-related OTUs along *Tara* Oceans stations, showing the relatively high abundance of the holobiont at Station 22. (**B**) Dorsal view of the entire acoel flatworm specimen (~300 μm). The epidermis (green) is completely covered with cilia and displays some pore holes. (**C**) The removal of the green channel reveals the widespread distribution of small unicellular algae (red areas) inside the acoel body. The worm's nuclei display a clear signal (compact round blue shapes), whereas the algal nuclei are dimmer. A dinoflagellate theca (arrowhead) is located in the central syncytium, likely indicating predation. (**D**) Cross-section along a *z*-*y* plane allows localization of the algae, beneath the epidermis in the parenchyma. Only the external cell layer (green signal) from the dorsal view is visible because of the thickness and opacity of the worm. Scale bar, 50 μm.

are available at www.raeslab.org/companion/ocean-interactome.html.

Random forest-based models

Eukaryotic, prokaryotic, and environmental matrices were merged into two matrices [deep chlorophyll maximum layer (DCM) and surface water layer (SRF)]. For each of the three models [OTU versus other OTUs (M_{OTU}), environmental factors (M_{ENV}) or combined ($M_{OTU+ENV}$)], regressions were perfomed with OTU abundance as dependent and the abundances of other OTUs or environmental factors as independent variables. For each regression, up to 20 independent variables were selected by using the minimum Redundancy Maximum Relevance (mRMR) filter-ranking algorithm. Random forest regression (*25*) was followed by a leave-one-out cross-validation. The variable subset with the minimum leave-one-out NMSE (normalized mean square error) was selected. To identify the best model for a given target OTU, the significance of the NMSE difference was tested on the absolute error values [paired Wilcoxon test adjusted by Benjamini-Hochberg false discovery rate (FDR) estimation (*74*)]. NMSE computed on random data are larger than those from original data. In addition, M_{ENV} outperformed M_{OTU} when OTU abundances were randomized.

Variance partitioning

Environmental variables were *z* score–transformed; spatial variables (MEM eigenvectors) were calculated based on latitude and longitude (*75*). Forward selection (*76*) was carried out with function forward.sel in R-package packfor. Significance of the selected variables was assessed with 1000 permutations by using functions rda and anova.cca in vegan. Variance partitioning (*77*) was performed by using function varpart in vegan on Hellinger-transformed abundance data, the forward-selected environmental variables, and the forward-selected spatial variables and tested for significance with 1000 permutations.

Network inference

Taxon-taxon co-occurrence networks were constructed as in (*78*), selecting Spearman and Kullback-Leibler dissimilarity measures. To compute *P* values, we first generated permutation and bootstrap distributions, with 1000 iterations each, by shuffling taxon abundances and resampling from samples with replacement, respectively. The measure-specific *P* value was then obtained as the probability of the null value (represented by the mean of the permutation distribution) under a Gauss curve fitted to the mean and standard deviation of the bootstrap distribution. Permutations computed for Spearman included a renormalization step, which mitigates compositionality bias (ReBoot). Measure-specific *P* values were merged by using Brown's method (*79*) and multiple-testing-corrected with Benjamini-Hochberg (*74*). Last, edges with an adjusted *P* value above 0.05, with a score below the thresholds (*30*) or not supported by both measures after assessment of significance, were discarded.

Taxon-environment networks were computed with the same procedure, starting with 8000 initial positive and negative edges, each supported by both methods. For computational efficiency, we computed 23 taxon-taxon and taxon-environment networks separately, for two depths (DCM and SRF), four eukaryotic size fractions (0.8 to 5 μm, >0.8 μm, 20 to 180 μm, and 180 to 2000 μm) and their combinations, the prokaryotic size fraction (0.2 to 1.6 μm and 0.2 to 3.0 μm) and its combination with each of the eukaryotic and virus (<0.2 μm) size fractions. We then generated 23 taxon-environment union networks for environmental triplet detection and merged the taxon-taxon networks into a global network with 92,633 edges.

Estimation of false discovery rate

We estimated the FDR of network construction with two null models. The first shuffles counts while preserving overall taxon proportions and total sample count sums, but removing any dependencies between taxa. For the second,

we fitted a Dirichlet-multinomial distribution to the input matrix using the dirmult package in R (*80*) and generated a null matrix by sampling from this distribution, preserving total sample count sums. Null matrices were generated from count matrices (0.8 to 5 μm, 20 to 180 μm, and 180 to 2000 μm eukaryotic and prokaryotic size fraction as well as bacteriophage-prokaryotic composite, SRF, and DCM). Network construction was performed with the 20 null matrices and thresholds applied to the original matrices (*28*). From edge numbers in the original and the null networks, we estimated an average FDR of 9% (*28*).

Indirect taxon edge detection

For each taxon-environment union network, node triplets consisting of two taxa and one environmental parameter were identified. For each triplet, interaction information II was computed as $II = CI(X, Y \mid Z) - I(X, Y)$, where *CI* is the conditional mutual information between taxa *X* and *Y* given environmental parameter *Z*, and *I* is the mutual information between *X* and *Y*. *CI* and *I* were estimated by using minet (*81*). Taxon edges in environmental triplets were considered indirect when $II < 0$ and within the 0.05 quantile of the random II distribution obtained by shuffling environmental vectors (500 iterations). If a taxon pair was part of more than one environmental triplet, the triplet with minimum interaction information was selected.

For each environmental triplet, we also checked whether its sign pattern (the combination of positive and/or negative correlations) was consistent with an indirect interaction. From eight possible patterns, four indicate indirect relationships (for example, two negatively correlated taxa correlated with opposite signs to an environmental factor).

Network deconvolution (*32*) was carried out with $\beta = 0.9$. We considered an environmental triplet as indirect according to network deconvolution if any of its edges were removed.

All (11,043) negative interaction information triplets were consistent with an indirect relationship according to their sign patterns, and a majority (8209) was also supported by network deconvolution.

Influence of ocean regions on co-occurrence patterns

Samples were divided into groups according to region membership. The impact of each sample group on the Spearman correlation of each edge in the network was assessed by dividing the (absolute) omission score (OS) (Spearman correlation without these samples) by the absolute original Spearman score. To account for group size, the OS was computed repeatedly for random, same-sized sample sets. Nonparametric *P* values were calculated as the number of times random OSs were smaller than the sample group OS, divided by number of random OS (500 for each taxon pair). Edges were classified as region-specific when the ratio of OS and absolute original score was below 1 and multiple-testing-corrected *P* values (Benjamini-Hochberg) were below 0.05.

Overrepresentation analysis

Significance of taxon–taxon counts at high taxonomic ranks was assessed with the hypergeometric distribution implemented in the R function phyper. Mutual exclusion versus copresence analysis was performed by using the binomial distribution implemented in the R function pbinom, with the background probability estimated by the frequency of edges in the network.

Oceanic region analysis was also assessed by use of R's pbinom function, with the background probability estimated by dividing total ocean-specific edge number by total edge number. The *P* value was computed as the probability of obtaining the observed number of ocean-specific edges among the edges of a taxon pair. The same procedure was repeated for each oceanic region separately, with region-specific success probabilities. Edges classified as indirect were discarded before the analysis.

In all tests, *P* values were adjusted for multiple testing according to Benjamini, Hochberg, and Yekutieli (BY), implemented in the R function p.adjust.

Extracting functional groups from the global plankton interactome

Functional groups consist of a mix of major monophyletic lineages of parasites, together with classical polyphyletic PFTs, as defined in (*10*, *54*, *55*). Metabarcodes in the network were sorted into 15 parasite groups and seven PFTs (*55*) according to their (i) taxonomical classification, (ii) membership in a given size fraction, (iii) trophic mode, and (iv) biogeochemical role in dimethyl sulfide (DMS) production or silicification. After mapping the metabarcodes and their edges onto PFTs and parasites, edges are weighted by the number of links they represent. Overrepresentation of the number of links included in each edge was assessed with the hypergeometric distribution.

Parasite links in large fractions may point to parasite-host connections. We extracted all edges in the large fractions (20 to 180 μm and 180 to 2000 μm) between barcodes annotated as parasites and nonparasitic barcodes. Partners of parasites comprised potential hosts (Fig. 3B) but also organisms that are either too small or without size information. The former may represent unknown parasites (for example, coinfecting a host with known parasites), whereas the latter may represent previously unknown hosts.

Nestedness and modularity analysis

The analysis was carried out for 1869 positively correlated phage-prokaryotic pairs. Modularity was computed with the LP (Label propagation) BRIM algorithm (*82*) in BiMAT (*83*) with 100 permutations. Nestedness of the host-phage network as quantified with the NODF (nestedness with overlap and decreasing fill) algorithm (*84*) in BiMAT with 100 permutations (preserving edge number and degree distribution) was significant, but not with the NTC algorithm (*85*). We also tested the impact of random removal or addition of 5, 10, 15, and 20% edges. After random addition/deletion of edges, modularity and nestedness (according to NODF) remained significant.

Confirmation of predicted viruses-host associations

Two different approaches were used to confirm virus-host associations predicted by the co-occurrence network. First, the network host prediction was compared with the "known" host for viral populations closely related to an isolated virus—populations with more than 50% of predicted genes affiliated to the same phage reference genome [based on a BLASTp against RefseqVirus, threshold of 10^{-03} on e-value and 50 on bit score (*18*)]. Known phages corresponded to viruses infecting SAR11, SAR116, and Cyanobacteria, so that a predicted host was considered correct if affiliated to Alphaproteobacteria, Alphaproteobacteria, and Cyanobacteria, respectively [the lowest rank for which there was taxonomic assignment for those bacterial OTUs (*69*)]. This procedure was repeated on 1000 randomized networks (with same-degree distribution) to calculate the significance of the results. Second, contigs of putative hosts predicted by co-occurrence analysis were compared with BLAST to a set of viral sequences detected in draft and single-cell genomes with VirSorter (https://pods.iplantcollaborative.org/wiki/display/DEapps/VIRSorter+1.0.2). One contig (36DCM_3902) (Fig. 3E) displayed significant sequence similarity (blastn *e*-value < 10^{-151} over two segments) to one contig detected in a single-cell genome (AA160P02DRAFT_ scaffold_31.32). In order to compare the putative host associated to each contig, rRNA genes were predicted in the single-cell amplified genome (SAG) contigs with meta-rRNA (*86*). Sequences were annotated based on BLAST against the nonredundant (nr) database, and the comparison plot was generated with Easyfig (*87*).

Literature-based evaluation of predicted protist interactions

A panel of four experts, two specialized in the study of planktonic mutualistic protists (C.d.V. and J.D.) and two specialized in the study of planktonic parasitic protists (C. Berney and N.H.), screened literature looking for symbiotic interactions occurring among eukaryotic plankton. From this search, they built a list of 574 known symbiotic interactions sensu lato (parasitism and mutualism, at least one protist partner) in marine eukaryotic plankton, covering 197 eukaryotic genera, described in 76 publications since 1971. The experts extracted only symbiotic interaction cases described either from direct observation of both interacting partners through microscope (45%), sequence from symbiont isolated from the observed host (14%), or both (41%). Direct observation of partners interacting (86%) provides high confidence for the interaction, and the symbiont sequence allows its taxonomic identification. The protocol to build the list was the following: (i) the experts manually screened 3170 publications associated to each PR2 db sequence http://ssu-rrna.org/pr2 (*73*); (ii) the experts screened 293 publications

retrieved from Web of Science with the following query: "TOPIC:(plankton* AND (marin* OR ocean*)) AND (parasit* OR symbios* OR mutualis*)"; (iii) the experts screened GenBank 18*S* rDNA sequences of symbionts for which the "host" field was known. They labeled these interactions as "Unpublished." Last, the experts discussed any observed discordance until agreement was reached. The final table of literature-curated interactions includes a column indicating the type of evidence gathered about the interaction: 1 for only getting symbiont sequence, 2 for direct observation, and 3 for both. Symbiont GenBank host field belongs to category 1.

Experimental validation of a predicted interaction

V9 pairs were searched for organisms of suitable size in order to allow its isolation from morphological samples. This way, we targeted a predicted photosymbiosis between an acoel flatworm [V9 rDNA metabarcode 83% similar to *Symsagittifera psammophila* (*88*)] and a photosynthetic microalga (*Tara* Oceans V9 metabarcode 100% similar to a *Tetraselmis* sp) (*89*).

Fifteen acoel specimens (hosts) were isolated from formaldehyde-4% microplankton samples of station 22 (A100000458), in which both partner OTUs displayed high abundances. Before imaging, specimens were rinsed with artificial seawater, then DNA and membrane structures were stained for 60 min with 10 µM Hoechst 33342 and 1.4 µM DiOC6(3) (Life Technologies, Grand Island, NY). Microscopy was conducted by using a Leica TCS SP8 (Leica Microsystems, Wetzlar, Germany) confocal laser scanning microscope and a HC PL APO 40x/1.10 W motCORR CS2 objective. The DiOC6 signal (ex488nm/em500-520nm) was collected simultaneously with the chlorophyll signal (ex488nm/em670-710nm), followed by the Hoechst signal (ex405 nm/em420-470nm). Images were processed with Fiji (*90*), and 3D specimens were reconstructed with Imaris (Bitplane, Belfast, UK).

To obtain the sequences of the metabarcodes of each partner, seven acoels were isolated from ethanol-preserved samples from station 22 (TARA_A100000451), individually rinsed in filtered seawater, and stored at –20°C in absolute ethanol. DNA was extracted with MasterPureTM DNA/RNA purification kit (Epicenter, Madison, WI) and polymerase chain reaction amplified by using the universal-eukaryote primers (forward 1389F and reverse 1510R) from (*10*). Chlorophyte-specific primers (Chloro2F: 5′- CGTATATTTAAGTTGYTGCAG-3′ and Tetra2-rev 5′- CAGCAATGGGCGGTGGC GAAC-3′) were designed to amplify the microalgae V9 rDNA as in (*4*). Purified amplicons were subjected to poly-A reaction and ligated in pCR®4-TOPO TA Cloning vector (Invitrogen, Carlsbad, CA), cloned by using chemically competent *Escherichia coli* cells, and Sanger-sequenced with the ABI-PRISM Big Dye Terminator Sequencing kit (Applied Biosystems, Foster City, CA) by using the 3130xl Genetic Analyzer (Applied Biosystems).

REFERENCES AND NOTES

1. F. Azam *et al.*, The ecological role of water-column microbes in the sea. *Mar. Ecol. Prog. Ser.* **10**, 257–263 (1983). doi: 10.3354/meps010257
2. A. W. Thompson *et al.*, Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. *Science* **337**, 1546–1550 (2012). doi: 10.1126/science.1222700; pmid: 22997339
3. A. Chambouvet, P. Morin, D. Marie, L. Guillou, Control of toxic marine dinoflagellate blooms by serial parasitic killers. *Science* **322**, 1254–1257 (2008). doi: 10.1126/science.1164387; pmid: 19023082
4. J. Decelle *et al.*, An original mode of symbiosis in open ocean plankton. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 18000–18005 (2012). doi: 10.1073/pnas.1212303109; pmid: 23071304
5. V. Smetacek, Making sense of ocean biota: How evolution and biodiversity of land organisms differ from that of the plankton. *J. Biosci.* **37**, 589–607 (2012). doi: 10.1007/s12038-012-9240-4; pmid: 22922185
6. J. L. Sabo, L. R. Gerber, "Trophic ecology," *AccessScience* (McGraw-Hill Education, 2014); available at www.accessscience.com/content/trophic-ecology/711650.
7. F. Rohwer, R. V. Thurber, Viruses manipulate the marine environment. *Nature* **459**, 207–212 (2009). doi: 10.1038/nature08060; pmid: 19444207
8. P. G. Verity, V. Smetacek, Organism life cycles, predation, and the structure of marine pelagic ecosystems. *Mar. Ecol. Prog. Ser.* **130**, 277–293 (1996). doi: 10.3354/meps130277
9. A. Z. Worden *et al.*, Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. *Science* **347**, 1257594 (2015). doi: 10.1126/science.1257594; pmid: 25678667
10. C. de Vargas *et al.*, Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).
11. K. Faust, J. Raes, Microbial interactions: From networks to models. *Nat. Rev. Microbiol.* **10**, 538–550 (2012). doi: 10.1038/nrmicro2832; pmid: 22796884
12. S. Chaffron, H. Rehrauer, J. Pernthaler, C. von Mering, A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res.* **20**, 947–959 (2010). doi: 10.1101/gr.104521.109; pmid: 20458099
13. J. Raes, I. Letunic, T. Yamada, L. J. Jensen, P. Bork, Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Mol. Syst. Biol.* **7**, 473 (2011). doi: 10.1038/msb.2011.6; pmid: 21407210
14. J. A. Gilbert *et al.*, Defining seasonal marine microbial community dynamics. *ISME J.* **6**, 298–308 (2012). doi: 10.1038/ismej.2011.107; pmid: 21850055
15. J. M. Beman, J. A. Steele, J. A. Fuhrman, Co-occurrence patterns for abundant marine archaeal and bacterial lineages in the deep chlorophyll maximum of coastal California. *ISME J.* **5**, 1077–1085 (2011). doi: 10.1038/ismej.2010.204; pmid: 21228895
16. C.-E. T. Chow, D. Y. Kim, R. Sachdeva, D. A. Caron, J. A. Fuhrman, Top-down controls on bacterial community structure: Microbial network analysis of bacteria, T4-like viruses and protists. *ISME J.* **8**, 816–829 (2014). doi: 10.1038/ismej.2013.199; pmid: 24196323
17. E. Karsenti *et al.*, A holistic approach to marine eco-systems biology. *PLOS Biol.* **9**, e1001177 (2011). doi: 10.1371/journal.pbio.1001177; pmid: 22028628
18. J. R. Brum *et al.*, Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498 (2015).
19. S. Sunagawa *et al.*, Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
20. E. Villar *et al.*, Environmental characteristics of Agulhas rings affect interocean plankton transport. *Science* **348**, 1261447 (2015).
21. Companion web site table w2; available at www.raeslab.org/companion/ocean_interactome/tables/W2.xlsx.
22. A. Meot, P. Legendre, D. Borcard, *Environ. Ecol. Stat.* **5**, 1–27 (1998). doi: 10.1023/A:1009693501830
23. Companion web site table w3; available at www.raeslab.org/companion/ocean_interactome/tables/W3.xlsx.
24. Companion web site figure w1; available at www.raeslab.org/companion/ocean_interactome/figures//W1.pdf .
25. L. Breiman, *Mach. Learn.* **45**, 5–32 (2001). doi: 10.1023/A:1010933404324
26. Companion web site table w4; available at www.raeslab.org/companion/ocean_interactome/tables/W4.xlsx.
27. Companion web site figure w2; available at http://www.raeslab.org/companion/ocean_interactome/figures/W2.pdf.
28. Companion web site table w5; available at www.raeslab.org/companion/ocean_interactome/tables/W5.xlsx.
29. Companion web site table w6; available at http://www.raeslab.org/companion/ocean_interactome/tables/W6.xlsx
30. Companion web site table w7; available at www.raeslab.org/companion/ocean_interactome/tables/W7.xlsx.
31. Companion web site figure w3; available at http://www.raeslab.org/companion/ocean_interactome/figures/W3.pdf.
32. S. Feizi, D. Marbach, M. Médard, M. Kellis, Network deconvolution as a general method to distinguish direct dependencies in networks. *Nat. Biotechnol.* **31**, 726–733 (2013). doi: 10.1038/nbt.2635; pmid: 23851448
33. Companion web site table w8 available at www.raeslab.org/companion/ocean_interactome/tables/W8.xlsx.
34. M. E. Cusick *et al.*, Literature-curated protein interaction datasets. *Nat. Methods* **6**, 39–46 (2009). doi: 10.1038/nmeth.1284; pmid: 19116613
35. Companion web site table w9; available at www.raeslab.org/companion/ocean_interactome/tables/W9.xlsx.
36. J. A. Fuhrman, J. A. Cram, D. M. Needham, Marine microbial community dynamics and their ecological interpretation. *Nat. Rev. Microbiol.* **13**, 133–146 (2015). doi: 10.1038/nrmicro3417; pmid: 25659323
37. Companion web site additional material is available at www.raeslab.org/companion/ocean_interactome/Accompanying_Material.docx.
38. Companion web site figure w4; available at www.raeslab.org/companion/ocean_interactome/figures/W4.pdf.
39. Companion web site table w10; available at www.raeslab.org/companion/ocean_interactome/tables/W10.xlsx.
40. G. B. McManus, L. F. Santoferrara, *The Biology and Ecology of Tintinnid Ciliates* (John Wiley & Sons, New York, 2012.
41. J. Cachon, in *Ann sci nat b*. (Paris, 1964), vol. 6, p. 1.
42. P. S. Salomon, E. Granéli, M. H. C. B. Neves, E. G. Rodríguez, Infection by Amoebophrya spp. parasitoids of dinoflagellates in a tropical marine coastal area. *Aquat. Microb. Ecol.* **55**, 143–153 (2009). doi: 10.3354/ame01293
43. F. Gómez, P. López-García, A. Nowaczyk, D. Moreira, The crustacean parasites Ellobiopsis Caullery, 1910 and Thalassomyces Niezabitowski, 1913 form a monophyletic divergent clade within the Alveolata. *Syst. Parasitol.* **74**, 65–74 (2009). doi: 10.1007/s11230-009-9199-1; pmid: 19633933
44. S. Ohtsuka *et al.*, Morphology and host-specificity of the apostome ciliate Vampyrophrya pelagica infecting pelagic copepods in the Seto Inland Sea, Japan. *Mar. Ecol. Prog. Ser.* **282**, 129–142 (2004). doi: 10.3354/meps282129
45. A. Skovgaard, S. A. Karpov, L. Guillou, The parasitic dinoflagellates Blastodinium spp. inhabiting the gut of marine, planktonic copepods: Morphology, ecology, and unrecognized species diversity. *Front. Microbiol.* **3**, 305 (2012). doi: 10.3389/fmicb.2012.00305; pmid: 22973263
46. L. Stemmann *et al.*, Global zoogeography of fragile macrozooplankton in the upper 100-1000 m inferred from the underwater video profiler. *ICES J. Mar. Sci.* **65**, 433–442 (2008). doi: 10.1093/icesjms/fsn010
47. J. M. Gasol, P. A. Del Giorgio, C. M. Duarte, *Biomass Distribution in Marine Planktonic Communities* (American Society of Limnology and Oceanography, Waco, TX, 1997), vol. 42.
48. B. A. Ward, S. Dutkiewicz, M. J. Follows, Modelling spatial and temporal patterns in size-structured marine plankton communities: Top-down and bottom-up controls. *J. Plankton Res.* **36**, 31–47 (2014). doi: 10.1093/plankt/fbt097
49. A. Ianora *et al.*, Aldehyde suppression of copepod recruitment in blooms of a ubiquitous planktonic diatom. *Nature* **429**, 403–407 (2004). doi: 10.1038/nature02526; pmid: 15164060
50. M. Martinez-Garcia *et al.*, Unveiling in situ interactions between marine protists and bacteria through single cell sequencing. *ISME J.* **6**, 703–707 (2012). doi: 10.1038/ismej.2011.126; pmid: 21938022
51. E. T. Jolley, A. K. Jones, The interaction between *Navicula muralis* grunow and an associated species of *Flavobacterium*. *Br. Phycol. J* **12**, 315–328 (1977). doi: 10.1080/00071617700650341
52. T. R. Miller, R. Belas, Dimethylsulfoniopropionate metabolism by Pfiesteria-associated Roseobacter spp. *Appl. Environ. Microbiol.* **70**, 3383–3391 (2004). doi: 10.1128/AEM.70.6.3383-3391.2004; pmid: 15184135
53. Companion web site figure w5; available at www.raeslab.org/companion/ocean_interactome/figures/W5.pdf.
54. C. Le Quere *et al.*, *Glob. Change Biol.* **11**, 17 (2005).
55. Companion web site table w11; available at www.raeslab.org/companion/ocean_interactome/tables/W11.xlsx.
56. A. Skovgaard, *Acta Protozool.* **53**, 51 (2014).
57. Companion web site figure w6; available at www.raeslab.org/companion/ocean_interactome/figures/W6.pdf.

58. S. von der Heyden, E. E. Chao, K. Vickerman, T. Cavalier-Smith, Ribosomal RNA phylogeny of bodonid and diplonemid flagellates and the evolution of euglenozoa. *J. Eukaryot. Microbiol.* **51**, 402–416 (2004). doi: 10.1111/j.1550-7408.2004.tb00387.x; pmid: 15352322
59. F. Gómez, D. Moreira, K. Benzerara, P. López-García, Solenicola setigera is the first characterized member of the abundant and cosmopolitan uncultured marine stramenopile group MAST-3. *Environ. Microbiol.* **13**, 193–202 (2011). doi: 10.1111/j.1462-2920.2010.02320.x; pmid: 20722698
60. C. E. Hamm *et al.*, Architecture and material properties of diatom shells provide effective mechanical protection. *Nature* **421**, 841–843 (2003). doi: 10.1038/nature01416; pmid: 12594512
61. P. Assmy, V. Smetacek, in *Encyclopedia of Microbiology*, M. Schaechter, Ed. (Elsevier, Oxford, UK, 2009), pp. 27–41.
62. Companion web site table w12; available at www.raeslab.org/companion/ocean_interactome/tables/W12.xlsx.
63. Companion web site figure w7; available at www.raeslab.org/companion/ocean_interactome/figures/W7.pdf.
64. J. S. Weitz *et al.*, Phage-bacteria infection networks. *Trends Microbiol.* **21**, 82–91 (2013). doi: 10.1016/j.tim.2012.11.003; pmid: 23245704
65. Companion web site table w13; available at www.raeslab.org/companion/ocean_interactome/tablesW13.xlsx.
66. Companion web site figure w9; available at www.raeslab.org/companion/ocean_interactome/companion_figures/W9.pdf.
67. S. Pesant *et al.*, Open science resources for the discovery and analysis of *Tara* Oceans data. http://biorxiv.org/content/early/2015/05/08/019117 (2015).
68. R. Logares *et al.*, Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ. Microbiol.* **16**, 2659–2671 (2013).
69. E. Pruesse *et al.*, SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35**, 7188–7196 (2007). doi: 10.1093/nar/gkm864; pmid: 17947321
70. R. C. Edgar, Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010). doi: 10.1093/bioinformatics/btq461; pmid: 20709691
71. Q. Wang, G. M. Garrity, J. M. Tiedje, J. R. Cole, Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007). doi: 10.1128/AEM.00062-07; pmid: 17586664
72. F. Mahé, T. Rognes, C. Quince, C. de Vargas, M. Dunthorn, Swarm: Robust and fast clustering method for amplicon-based studies. *PeerJ* **2**, e593 (2014). doi: 10.7717/peerj.593; pmid: 25276506
73. L. Guillou *et al.*, The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* **41** (D1), D597–D604 (2013). doi: 10.1093/nar/gks1160; pmid: 23193267
74. Y. Benjamini, Y. Hochberg, *J. R. Stat. Soc., B* **57**, 289 (1995).
75. D. Borcard, P. Legendre, C. Avois-Jacquet, H. Tuomisto, DIssecting the spatial structure of ecological data at multiple scales. *Ecology* **85**, 1826–1832 (2004). doi: 10.1890/03-3111
76. F. G. Blanchet, P. Legendre, D. Borcard, Forward selection of explanatory variables. *Ecology* **89**, 2623–2632 (2008). doi: 10.1890/07-0986.1; pmid: 18831183
77. D. Borcard, P. Legendre, P. Drapeau, Partialling out the spatial component of ecological variation. *Ecology* **73**, 1045 (1992). doi: 10.2307/1940179
78. K. Faust *et al.*, Microbial co-occurrence relationships in the human microbiome. *PLOS Comput. Biol.* **8**, e1002606 (2012). doi: 10.1371/journal.pcbi.1002606; pmid: 22807668
79. M. B. Brown, 400: A method for combining non-independent, one-sided tests of significance. *Biometrics* **31**, 987 (1975). doi: 10.2307/2529826
80. T. Tvedebrink, Overdispersion in allelic counts and θ-correction in forensic genetics. *Theor. Popul. Biol.* **78**, 200–210 (2010). doi: 10.1016/j.tpb.2010.07.002; pmid: 20633572
81. P. E. Meyer, F. Lafitte, G. Bontempi, minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics* **9**, 461 (2008). doi: 10.1186/1471-2105-9-461; pmid: 18959772
82. L. Xin, T. Murata, in *Web Intelligence and Intelligent Agent Technologies, 2009* (WI-IAT 09. IEEE/WIC/ACM International Joint Conferences, 2009), vol. 1, pp. 50.
83. C. O. Flores, T. Poisot, S. Valverde, J. S. Weitz, http://arxiv.org/abs/1406.6732 (2014)
84. M. J. Barber, Modularity and community detection in bipartite networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **76**, 066102 (2007). doi: 10.1103/PhysRevE.76.066102; pmid: 18233893
85. W. Atmar, B. D. Patterson, The measure of order and disorder in the distribution of species in fragmented habitat. *Oecologia* **96**, 373–382 (1993). doi: 10.1007/BF00317508
86. Y. Huang, P. Gilna, W. Li, Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics* **25**, 1338–1340 (2009). doi: 10.1093/bioinformatics/btp161; pmid: 19346323
87. M. J. Sullivan, N. K. Petty, S. A. Beatson, Easyfig: A genome comparison visualizer. *Bioinformatics* **27**, 1009–1010 (2011). doi: 10.1093/bioinformatics/btr039; pmid: 21278367
88. I. Ruiz-Trillo, M. Riutort, D. T. J. Littlewood, E. A. Herniou, J. Baguña, Acoel flatworms: Earliest extant bilaterian Metazoans, not members of Platyhelminthes. *Science* **283**, 1919–1923 (1999). doi: 10.1126/science.283.5409.1919; pmid: 10082465
89. R. J. Gast, T. A. McDonnell, D. A. Caron, srDna-based taxonomic affinities of algal symbionts from a planktonic foraminifer and a solitary radiolarian. *J. Phycol.* **36**, 172–177 (2000). doi: 10.1046/j.1529-8817.2000.99133.x
90. J. Schindelin *et al.*, Fiji: An open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012). doi: 10.1038/nmeth.2019; pmid: 22743772
91. M. E. Newman, Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **74**, 036104 (2006). doi: 10.1103/PhysRevE.74.036104; pmid: 17025705

ACKNOWLEDGMENTS

We thank the commitment of the following people and sponsors: Centre National de la Recherche Scientifique (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 (G.L.M., K.F., S.C., and J.R.), Rega Institute (J.R.), KU Leuven (J.R.), The French Ministry of Research, the French Government "Investissements d'Avenir" programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), MEMO LIFE (ANR-10-LABX-54), PSL* Research University (ANR-11-IDEX-0001-02), ANR (projects POSEIDON/ANR-09-BLAN-0348, PHYTBACK/ANR-2010-1709-01, PROMETHEUS/ANR-09-PCS-GENM-217, TARA GIRUS/ANR-09-PCS-GENM-218, European Union FP7 (MicroB3/No.287589, IHMS/HEALTH-F4-2010-261376, ERC Advanced Grant Awards to CB (Diatomite: 294823), Gordon and Betty Moore Foundation grant (3790) to M.B.S., Spanish Ministry of Science and Innovation grant CGL2011-26848/BOS MicroOcean PANGENOMICS to S.G.A., TANIT (CONES 2010-0036) from the Agència de Gestió d´Ajusts Universitaris i Reserca funded to S.G.A., JSPS KAKENHI grant number 26430184 to H.O., FWO, BIO5, Biosphere 2, Agnès b., the Veolia Environment Foundation, Region Bretagne, Lorient Agglomeration, World Courier, Illumina, the EDF Foundation, FRB, the Prince Albert II de Monaco Foundation, Etienne Bourgois, the *Tara* schooner, and its captain and crew. We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries that graciously granted sampling permissions. *Tara* Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org). We also acknowledge the EMBL Advanced Light Microscopy Facility (ALMF), and in particular R. Pepperkok. The authors further declare that all data reported herein are fully and freely available from the date of publication, with no restrictions, and that all of the samples, analyses, publications, and ownership of data are free from legal entanglement or restriction of any sort by the various nations whose waters the *Tara* Oceans expedition sampled in. Data described herein is available at www.raeslab.org/companion/ocean-interactome.html, at the EBI under the projects PRJEB402 and PRJEB6610, and at Pangaea http://doi.pangaea.de/10.1594/PANGAEA.840721, http://doi.pangaea.de/10.1594/PANGAEA.840718, http://doi.pangaea.de/10.1594/PANGAEA.843018 and http://doi.pangaea.de/10.1594/PANGAEA.843022 and on table S1. The data release policy regarding future public release of *Tara* Oceans data are described in Pesant *et al.* (67). All authors approved the final manuscript. This article is contribution number 25 of Tara Oceans. The supplementary materials contain additional data.

TARA OCEANS COORDINATORS

Silvia G. Acinas,[1] Peer Bork,[2] Emmanuel Boss,[3] Chris Bowler,[4] Colomban De Vargas,[5,6] Michael Follows,[7] Gabriel Gorsky,[8,9] Nigel Grimsley,[10,11] Pascal Hingamp,[12] Daniele Iudicone,[13] Olivier Jaillon,[14,15,16] Stefanie Kandels-Lewis,[2] Lee Karp-Boss,[3] Eric Karsenti,[17,18] Uros Krzic,[19] Fabrice Not,[5,6] Hiroyuki Ogata,[20] Stephane Pesant,[21,22] Jeroen Raes,[23,24,25] Emmanuel G. Reynaud,[26] Christian Sardet,[8] Mike Sieracki,[27] Sabrina Speich,[28,29] Lars Stemmann,[8] Matthew B. Sullivan,[30] Shinichi Sunagawa,[2] Didier Velayoudon,[31] Jean Weissenbach,[14,15,16] Patrick Wincker[14,15,16]

[1]Department of Marine Biology and Oceanography, Institute of Marine Sciences ICM-CSIC, Barcelona, Spain. [2]Structural and Computational Biology, European Molecular Biology Laboratory, Heidelberg, Germany. [3]School of Marine Sciences, University of Maine, Orono, USA. [4]Environmental and Evolutionary Genomics Section, Institut de Biologie de l'Ecole Normale Supérieure, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 8197, Institut National de la Santé et de la Recherche Médicale U1024, Ecole Normale Supérieure, Paris, France. [5]CNRS, UMR 7144, Station Biologique de Roscoff, Roscoff, France. [6]Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, Roscoff, France. [7]Dept of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, USA. [8]CNRS, UMR 7093, Laboratoire d'Océanographie de Villefranche (LOV), Observatoire Océanologique, F-06230 Villefranche-sur-mer, France. [9]Sorbonne Universités, UPMC Paris 06, UMR 7093, Laboratoire d'Océanographie de Villefranche (LOV), Observatoire Océanologique, F-06230 Villefranche-sur-mer, France. [10]CNRS UMR 7232, BIOM, Banyuls-sur-Mer, France. [11]Sorbonne Universités, OOB, UPMC Paris 06, Banyuls-sur-Mer, France. [12]Aix Marseille Université, CNRS, IGS UMR 7256, Marseille, France. [13]Laboratory of Ecology and Evolution of Plankton, Stazione Zoologica Anton Dohrn, Naples, Italy. [14]CEA, Genoscope, Evry France. [15]CNRS, UMR 8030, Evry, France. [16]Université d'Evry, UMR 8030, Evry, France. [17]Environmental and Evolutionary Genomics Section, Institut de Biologie de l'Ecole Normale Supérieure, CNRS, UMR 8197, Institut National de la Santé et de la Recherche Médicale U1024, Ecole Normale Supérieure, Paris, France. [18]Directors' Research, European Molecular Biology Laboratory, Heidelberg, Germany. [19]Cell Biology and Biophysics, European Molecular Biology Laboratory, Heidelberg, Germany. [20]Institute for Chemical Research, Kyoto University, Kyoto, Japan. [21]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany. [22]MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany. [23]Department of Microbiology and Immunology, Rega Institute KU Leuven, Leuven, Belgium. [24]VIB Center for the Biology of Disease, VIB, Leuven, Belgium. [25]Laboratory of Microbiology, Vrije Universiteit Brussel, Brussels, Belgium. [26]School of Biology and Environmental Science, University College Dublin, Dublin, Ireland. [27]Bigelow Laboratory for Ocean Sciences, East Boothbay, USA. [28]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, Paris, France. [29]Laboratoire de Physique des Océan, UBO-IUEM, Polouzané, France. [30]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, USA. [31]DVIP Consulting, Sèvres, France.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/1262073/suppl/DC1
Table S1

3 October 2014; accepted 18 March 2015
10.1126/science.1262073

E. Co-authored manuscript 3: Villar et al, 2015

Villar et al., 2015:

I was in charge of computing the diversity indexes (Shannon, Richness and Simpson) for the top 6 eukaryotic groups (Bacillariophyta, Collodaria, Copepoda, Dinoflagellates, MALV and Haptophytes), in fractions 0.8-5, 20-180, and 180-2000 microns.

To normalize for differences in sequencing effort, V9 rDNA barcode libraries were resampled 50 times for the number of reads corresponding to the smallest library in each size fraction (provided by E.Villar): 0.8 to 5 microns, 776,358 reads; 20 to 180 microns, 1,170,592 reads; and 180 to 2000 microns, 767,940 reads. V9 rDNA barcode counts were then converted to the average number of times seen in the 50 resampling events, and barcodes with less than 10 reads were removed as potential sequencing artifacts. We used downsampled barcode richness (number of distinct V9 rDNA barcodes) as a diversity descriptor because using V9 rDNA barcode abundances to compare plankton assemblages would likely be biased.

The figure was finally replaced by Table S4 of the manuscript.

OCEAN PLANKTON

Environmental characteristics of Agulhas rings affect interocean plankton transport

Emilie Villar,[1*] Gregory K. Farrant,[2,3†] Michael Follows,[11†] Laurence Garczarek,[2,3†] Sabrina Speich,[5,23¶] Stéphane Audic,[2,3] Lucie Bittner,[2,3,4‡] Bruno Blanke,[5] Jennifer R. Brum,[6**] Christophe Brunet,[7] Raffaella Casotti,[7] Alison Chase,[8] John R. Dolan,[9,10] Fabrizio d'Ortenzio,[9,10] Jean-Pierre Gattuso,[9,10] Nicolas Grima,[5] Lionel Guidi,[9,10] Christopher N. Hill,[11] Oliver Jahn,[11] Jean-Louis Jamet,[12] Hervé Le Goff,[13] Cyrille Lepoivre,[1] Shruti Malviya,[4] Eric Pelletier,[14,15,16] Jean-Baptiste Romagnan,[9,10] Simon Roux,[6**] Sébastien Santini,[1] Eleonora Scalco,[7] Sarah M. Schwenck,[6] Atsuko Tanaka,[4§] Pierre Testor,[13] Thomas Vannier,[14,15,16] Flora Vincent,[4] Adriana Zingone,[7] Céline Dimier,[2,3,4] Marc Picheral,[9,10] Sarah Searson,[9,10||] Stefanie Kandels-Lewis,[17,18] *Tara* Oceans Coordinators¶ Silvia G. Acinas,[19] Peer Bork,[17,20] Emmanuel Boss,[8] Colomban de Vargas,[2,3] Gabriel Gorsky,[9,10] Hiroyuki Ogata,[1#] Stéphane Pesant,[21,22] Matthew B. Sullivan,[6**] Shinichi Sunagawa,[17] Patrick Wincker,[14,15,16] Eric Karsenti,[4,18*] Chris Bowler,[4*] Fabrice Not,[2,3*††] Pascal Hingamp,[1*††] Daniele Iudicone[7*††]

Agulhas rings provide the principal route for ocean waters to circulate from the Indo-Pacific to the Atlantic basin. Their influence on global ocean circulation is well known, but their role in plankton transport is largely unexplored. We show that, although the coarse taxonomic structure of plankton communities is continuous across the Agulhas choke point, South Atlantic plankton diversity is altered compared with Indian Ocean source populations. Modeling and in situ sampling of a young Agulhas ring indicate that strong vertical mixing drives complex nitrogen cycling, shaping community metabolism and biogeochemical signatures as the ring and associated plankton transit westward. The peculiar local environment inside Agulhas rings may provide a selective mechanism contributing to the limited dispersal of Indian Ocean plankton populations into the Atlantic.

The Agulhas Current, which flows down the east coast of Africa, leaks from the Indo-Pacific Ocean into the Atlantic Ocean (*1*). This leakage, a choke point to heat and salt distribution across the world's oceans, has been increasing over the last decades (*2*). The influence of the Agulhas leakage on global oceanic circulation makes this area a sensitive lever in climate change scenarios (*3*). Agulhas leakage has been a gateway for planetary-scale water transport since the early Pleistocene (*4*), but diatom fossil records suggest that it is not a barrier to plankton dispersal (*5*). Most of the Agulhas leakage occurs through huge anticyclonic eddies known as Agulhas rings. These 100- to 400-km-diameter rings bud from Indian Ocean subtropical waters at the Agulhas Retroflection (*1*). Each year, up to half a dozen Agulhas rings escape the Indian Ocean, enter Cape Basin, and drift northwesterly across the South Atlantic, reaching the South American continent over the course of several years (*1*, *6*). During the transit of Agulhas rings, strong westerly "roaring forties" winds prevalent in the southern 40s and 50s latitudes cause intense internal cooling and mixing (*7*).

We studied the effect of Agulhas rings and the environmental changes they sustain on plankton dispersal. Plankton such as microalgae, which produce half of the atmospheric oxygen derived from photosynthesis each year, are at the base of open-ocean ecosystem food chains, thus playing an essential role in the functioning of the biosphere. Their dispersal is critical for marine ecosystem resilience in the face of environmental change (*8*). As part of the *Tara* Oceans expedition (*9*), we describe taxonomic and functional plankton assemblages inside Agulhas rings and across the three oceanic systems that converge at the Agulhas choke point: the western Indian Ocean subtropical gyre, the South Atlantic Ocean gyre, and the Southern Ocean below the Antarctic Circumpolar Current (Fig. 1).

Physical and biological oceanography of the sampling sites

The Indian, South Atlantic, and Southern Oceans were each represented by three sites sampled between May 2010 and January 2011 (Fig. 1 and table S1). A wide range of environmental conditions were encountered (*10*). We first sampled the two large contiguous Indian and South Atlantic subtropical gyres and the Agulhas ring structures that maintain the physical connection between them. On the western side of the Indian Ocean, station TARA_052 was characterized by tropical, oligotrophic conditions. Station TARA_064 was located within an anticyclonic eddy representing the Agulhas Current recirculation. Station TARA_065 was located at the inner edge of the Agulhas Current on the South African slope that feeds the Agulhas retroflection and Agulhas ring formation (*3*). In the South Atlantic Ocean, station TARA_070, sampled in late winter, was located in the eastern subtropical Atlantic basin. Station TARA_072 was located within the tropical circulation of the South Atlantic Ocean, and Station TARA_076 was at the northwest extreme of the South Atlantic subtropical gyre. Two stations (TARA_068 and TARA_078) from the west and east South Atlantic Ocean sampled Agulhas rings. Three stations (TARA_082, TARA_084, and TARA_085) in the Southern Ocean were selected to sample the Antarctic Circumpolar Current frontal system. Station TARA_082 sampled sub-Antarctic waters flowing northward along the Argentinian slope, waters that flow along the Antarctic Circumpolar Current (*11*) with characteristics typical

[1]Aix Marseille Université, CNRS, IGS UMR 7256, 13288 Marseille, France. [2]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [3]Sorbonne Universités, Université Pierre et Marie Curie UPMC, Université Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [4]Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), and Inserm U1024, and CNRS UMR 8197, F-75005 Paris, France. [5]Laboratoire de Physique des Océans (LPO) UMR 6523 CNRS-Ifremer-IRD-UBO, Plouzané, France. [6]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA. [7]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [8]School of Marine Sciences, University of Maine, Orono, ME, USA. [9]Sorbonne Universités, UPMC Université Paris 06, Observatoire Océanologique, F-06230 Villefranche-sur-Mer, France. [10]INSU-CNRS, UMR 7093, LOV, Observatoire Océanologique, F-06230 Villefranche-sur-Mer, France. [11]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [12]Université de Toulon, Laboratoire PROTEE-EBMA E.A. 3819, BP 20132, 83957 La Garde Cedex, France. [13]CNRS, UMR 7159, Laboratoire d'Océanographie et du Climat LOCEAN, 4 Place Jussieu, 75005 Paris, France. [14]Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), Institut de Génomique, Genoscope, 2 Rue Gaston Crémieux, 91057 Evry, France. [15]CNRS, UMR 8030, CP5706, Evry, France. [16]Université d'Evry, UMR 8030, CP5706, Evry, France. [17]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [18]Directors' Research, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [19]Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM), CSIC, Passeig Marítim de la Barceloneta, 37-49, Barcelona E08003, Spain. [20]Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany. [21]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany. [22]MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany. [23]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD) UMR 8539, Ecole Normale Supérieure, 24 Rue Lhomond, 75231 Paris Cedex 05, France. ***Corresponding author. E-mail: villar@igs.cnrs-mrs.fr (E.V.); not@sb-roscoff.fr (F.N.); hingamp@igs.cnrs-mrs.fr (P.H.); iudicone@szn.it (D.I.); karsenti@embl.de (E.K.), cbowler@biologie.ens.fr (C.B.)** †These authors contributed equally to this work. ‡Present address: CNRS FR3631, Institut de Biologie Paris-Seine, F-75005 Paris, France; Sorbonne Universités, UPMC Université Paris 06, Institut de Biologie Paris-Seine (IBPS), F-75005 Paris, France. §Present address: Muroran Marine Station, Field Science Center for Northern Biosphere, Hokkaido University, Japan. ||Present address: CMORE, University Hawaii, Honolulu, USA. ¶*Tara* Oceans coordinators are listed at the end of this paper. #Present address: Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan. **Present address: Department of Microbiology, Ohio State University, Columbus, OH 43210, USA. ††These authors contributed equally to this work.

of summer sub-Antarctic surface waters and are stratified by seasonal heating. Station TARA_084 was located on the southern part of the Antarctic Circumpolar Current, in the Drake Passage between the Polar Front and the South Antarctic Circumpolar Current front (*11*). Station TARA_085 was located on the southern edge of the South Antarctic Circumpolar Current front with waters typical of polar regions.

We compared overall plankton community structures between the three oceans using imaging and genetic surveys of samples from the epipelagic zone of each station (*12*). Prokaryote, phyto-, and zooplankton assemblages were similar across Indian and South Atlantic Ocean samples but different from Southern Ocean samples (Fig. 2A). In the Indian and South Atlantic Oceans, zooplankton communities were dominated by Calanoida, Cyclopoida (Oithonidae), and Poecilostomatoida copepods (*12*); phytoplankton communities were mainly composed of chlorophytes, pelagophytes, and haptophytes (*12*). In contrast, Southern Ocean zooplankton communities were distinguished by an abundance of *Limacina* spp. gastropods and Poecilostomatoida copepods. Southern Ocean phytoplankton were primarily diatoms and haptophytes. The divergence was even more conspicuous with respect to prokaryotes, in that picocyanobacteria, dominant in the Indian and South Atlantic Oceans, were absent in the Southern Ocean. The Southern Ocean had a high proportion of Flavobacteria and Rhodobacterales (*12*). Virus concentrations in the <0.2-μm size fractions were significantly lower in the southernmost Southern Ocean station (*13*). Viral particles were significantly smaller in two of the three Southern Ocean sampling sites, and two Southern Ocean viromes had significantly lower richness compared with the South Atlantic and Indian Oceans (*13*). Although nucleocytoplasmic large DNA viruses were similarly distributed in the South Atlantic and Indian Oceans (*12*), two Southern Ocean sites contained coccolithoviruses also found in the TARA_068 Agulhas ring but not in the other Indian and South Atlantic stations.

Biological connection across the Agulhas choke point

Genetic material as represented by ribosomal RNA gene (rDNA) sequences showed exchange patterns across the oceans (shared barcode richness) (*14*). Despite a smaller interface between the Indian and South Atlantic Oceans than either have with the Southern Ocean, more than three times as much genetic material was in common between the Indian and South Atlantic Oceans than either had with the Southern Ocean (Fig. 2B) (*15*). Indeed, the Indian–South Atlantic interocean shared barcodes richness (32 ± 5%) was not significantly different from typical intraocean values (37 ± 7%, Tukey post hoc, 0.95 confidence). Shared barcode richness involving the Southern Ocean was significantly lower (9 ± 3%) (Fig. 2C). We found that the proportion of whole shotgun metagenomic reads shared between samples, both intraoceanic and Indian–South Atlantic interocean similarities, were in the 18 to 30% range, whereas interocean similarities with Southern Ocean samples were only 5 to 6% (*16*). The statistically indistinguishable Indo-Atlantic intra- and interocean genetic similarities revealed a high Indo-Atlantic biological connection despite the physical basin discontinuity.

Nonetheless, differences on either side of the Agulhas choke point were evident. We found that prokaryote barcode richness was greater in the South Atlantic than in the Indian Ocean (Fig. 3A) (0.2- to 3-μm size fraction). The opposite trend characterized eukaryotes larger than 20 μm in size. We cannot rule out the possibility that the higher prokaryote diversity observed in the South Atlantic Ocean might be due to a protocol artifact resulting from a difference in prefiltration pore size from 1.6 μm (Indian Ocean) to 3 μm (South Atlantic and Southern Oceans). As also evident from the panoceanic *Tara* Oceans data set (*17*), smaller size fractions showed greater eukaryote diversity across the Agulhas system. In all size fractions that we analyzed, samples from the Southern Ocean were less diverse than samples from the South Atlantic Ocean and Indian Ocean (Fig. 3A).

When rDNA barcodes were clustered by sequence similarity and considered at operational taxonomic unit (OTU) level (*14*), more than half (57%) of the OTUs contained higher sub-OTU barcode richness in the Indian Ocean than in the South Atlantic Ocean, whereas less than a third (32%) of OTUs were richer in the South Atlantic Ocean, leaving only 11% as strictly cosmopolitan (Fig. 3B). Taken together, these 1307 OTUs represented 98% of the barcode abundance, indicating that the observed higher barcode richness within OTUs in the Indian Ocean was not conferred by the rare biosphere. Certain taxa displayed unusual sub-OTU richness profiles across the choke point. Consistent with their relatively large size, Opisthokonta (mostly copepods), Rhizaria (such as radiolarians), and Stramenopiles (in particular diatoms) had much higher sub-OTU barcode richness in the Indian Ocean, whereas only small-sized Hacrobia (mostly haptophytes) showed modest increased sub-OTU barcode richness in the South Atlantic Ocean. The plankton filtering that we observed in fractions above 20 μm through the Agulhas choke point might explain the reduction of marine nekton diversity from the Indian Ocean to the South Atlantic Ocean (*18*) by propagating up the food web (*19*).

In situ sampling of two Agulhas rings

To understand whether the environment of Agulhas rings, the main transporters of water across the choke point, might act as a biological filter between the Indian Ocean and the South Atlantic Ocean, we analyzed data collected in both a young and an old Agulhas ring. The young ring sampled at station TARA_068 was located in the Cape Basin, west of South Africa, where rings are often observed after their formation at the Agulhas Retroflection (*7*, *20*). It was a large Agulhas ring that detached from the retroflection about 9 to 10 months before sampling. This ring first moved northward and then westward in the Cape Basin while interacting with other structures (red track in Fig. 1) (*21*). Ocean color data collected by satellite showed that surface chlorophyll concentrations were higher in the Cape Basin than at the retroflection, suggesting that vigorous vertical

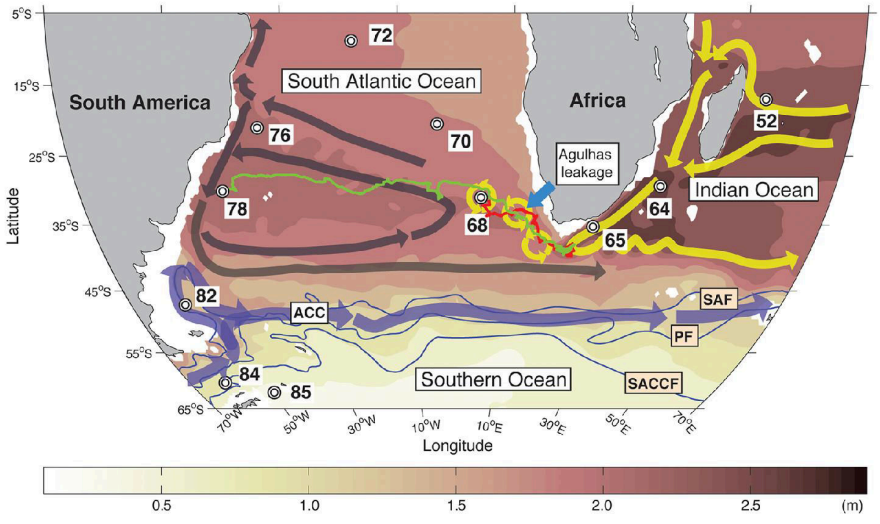


Fig. 1. The oceanic circulation around the Agulhas choke point and location of *Tara* Oceans stations. The map shows the location of sampling stations, together with trajectories of the young and old Agulhas rings (TARA_068 and TARA_078, red and green tracks, respectively). The stations here considered as representative of the main basins are (i) TARA_052, TARA_064, and TARA_065 for Indian Ocean; (ii) TARA_070, TARA_072, and TARA_076 for the South Atlantic Ocean, and (iii) TARA_082, TARA_084, and TARA_085 for the Southern Ocean. The mean ocean circulation is schematized by arrows (currents) and background colors [surface climatological dynamic height (0/2000 dbar from CARS2009; www.cmar.csiro.au/cars)] (*70*). Agulhas rings are depicted as circles. The Antarctic Circumpolar Current front positions are from (*13*).

mixing might have occurred in the Cape Basin (*22*). At the time of sampling, the anticyclonic Agulhas ring was 130 to 150 km in diameter, was about 30 cm higher than average sea surface height, and was flanked by a 130- to 150-km cyclonic eddy to the north and a larger (>200 km) one to the east (Fig. 4A) (*23*). Thermosalinograph data showed that filaments of colder, fresher water surrounded the young ring core (Fig. 4A) (*23*). To position the biological sampling station close to the ring core, a series of conductivity-temperature-depth (CTD) casts was performed (*23*, *24*). The young Agulhas ring had a surface temperature and salinity of 16.8°C and 35.7 practical salinity units (PSU), respectively, and the isopycnal sloping could be traced down to CTD maximal depth (900 to 1000 m). The core of the ring water was 5°C cooler than Indian Ocean subtropical source waters at similar latitudes (TARA_065) (table S1), typical for the subtropical waters south of Africa (17.8°C, 35.56 PSU, respectively) (*25*). The mixed layer of the young ring was deep (>250 m) compared with seasonal cycles of the mixed layer depths in the region (50 to 100 m) (Fig. 4C), typical of Agulhas rings (*26*). At larger scales (Fig. 4B) (*24*), steep spatial gradients were observed, with fresher and colder water in the Cape Basin than in the Agulhas Current because of both lateral mixing with waters from the south and surface fluxes. This confirms that the low temperature of the young Agulhas ring is a general feature of this Indian to South Atlantic Ocean transitional basin. Air-sea exchanges of heat and momentum promoted convection in the ring core, which was not compensated by lateral mixing and advection. The core of the Agulhas ring thus behaved as a subpolar environment traveling across a subtropical region.

At station TARA_078, we sampled a second structure whose origins were in the Agulhas Retroflection, likely a 3-year-old Agulhas ring. This old ring, having crossed the South Atlantic Ocean, was being absorbed by the western boundary current of the South Atlantic subtropical gyre. The structure sampled at station TARA_078 was characterized by a warm salty core (*27*). As for the young Agulhas ring sampled, the old ring also had a 100-m-deeper pycnocline than surrounding waters, typical of large anticyclonic structures.

The plankton assemblage of both Agulhas rings most closely resembled the assemblages found in Indian and South Atlantic samples (Fig. 2A). At higher resolution, barcodes (Fig. 2, B and C) and metagenomic reads (*16*) shared between the Agulhas rings and the Indian or South Atlantic samples showed that the young ring was genetically distinct from both Indian and South Atlantic samples,

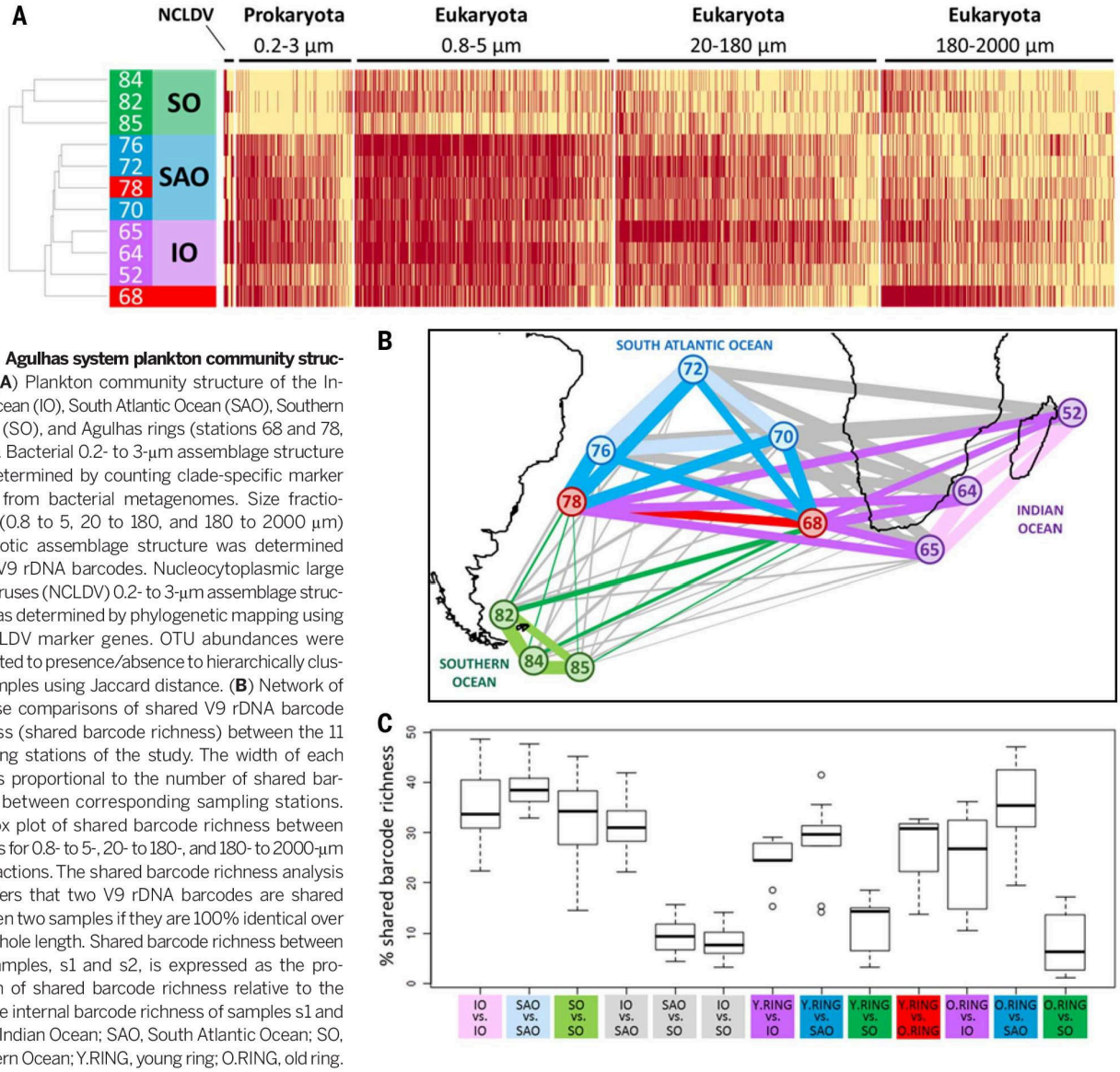


Fig. 2. Agulhas system plankton community structure. (**A**) Plankton community structure of the Indian Ocean (IO), South Atlantic Ocean (SAO), Southern Ocean (SO), and Agulhas rings (stations 68 and 78, in red). Bacterial 0.2- to 3-µm assemblage structure was determined by counting clade-specific marker genes from bacterial metagenomes. Size fractionated (0.8 to 5, 20 to 180, and 180 to 2000 µm) eukaryotic assemblage structure was determined using V9 rDNA barcodes. Nucleocytoplasmic large DNA viruses (NCLDV) 0.2- to 3-µm assemblage structure was determined by phylogenetic mapping using 16 NCLDV marker genes. OTU abundances were converted to presence/absence to hierarchically cluster samples using Jaccard distance. (**B**) Network of pairwise comparisons of shared V9 rDNA barcode richness (shared barcode richness) between the 11 sampling stations of the study. The width of each edge is proportional to the number of shared barcodes between corresponding sampling stations. (**C**) Box plot of shared barcode richness between stations for 0.8- to 5-, 20- to 180-, and 180- to 2000-µm size fractions. The shared barcode richness analysis considers that two V9 rDNA barcodes are shared between two samples if they are 100% identical over their whole length. Shared barcode richness between two samples, s1 and s2, is expressed as the proportion of shared barcode richness relative to the average internal barcode richness of samples s1 and s2. IO, Indian Ocean; SAO, South Atlantic Ocean; SO, Southern Ocean; Y.RING, young ring; O.RING, old ring.

whereas the old ring was similar to its surrounding South Atlantic samples (Tukey post hoc, 0.95 confidence). Light microscopy analyses revealed some plankton groups specific to the young Agulhas ring, such as *Pseudo-nitzschia* spp., which represented 20% of the phytoplankton counts but less than 10% in all other stations (*12*). Other potentially circumstantial plankton characteristic of the young Agulhas ring included the tintinnid *Dictyocysta pacifica* (*12*), the diatom *Corethron pennatum* (*12*), and the dinoflagellate *Tripos limulus* (*12*). A tiny (less than 15 µm long) pennate diatom from the genus *Nanoneis*, which we saw only in the young Agulhas ring and Indian Ocean stations around the African coasts (*28*), was an example of the Indo-Atlantic plankton diversity filtering observed at rDNA barcode level and corroborated by microscopy. OTU clustered barcodes revealed a variety of young Agulhas ring sub-OTU richness patterns compared with source and destination oceans (Fig. 5A). Among Copepoda, *Gaetanus variabilis* and *Corycaeus speciosus* were the more cosmopolitan species (Fig. 5B), whereas *Bradya* species found in the young ring were mainly similar to those from the Indian Ocean. *Acartia negligens* and *Neocalanus robustior* displayed high levels of barcode richness specific to each side of the Agulhas choke point. Bacillariophyceae were heavily filtered from Indian to South Atlantic Oceans (Fig. 5C), and most OTUs (17 out of 20) were absent in the young ring, suggesting that diversity filtering could take place earlier in the ring's 9-month history. Consistent with the observed particularities of the plankton in the young ring, continuous underway optical measurements showed that the ring core photosynthetic community differed from surrounding waters (*29*–*31*). Intermediate size cells, and relatively low content of photoprotective pigments, reflected low growth irradiance and suggested a transitional physiological state. Thus, the plankton community in the young Agulhas ring had diverged from plankton communities typical of its original Indian waters but, even 9 months after formation, had not converged with its surrounding South Atlantic waters.

Deep mixing in Agulhas rings promotes plankton bloom

The upper water column of the young ring showed a high nitrite concentration (>0.5 mmol m^{-3}) (Fig. 4D) (*32*). This observation, along with its particularly deep mixed layer (>250 m), suggested that as Agulhas rings proceed westward in the Cape Basin, vigorous deep mixing of their weakly stratified waters may have entrained nitrate and stimulated phytoplankton blooms. Typically, fresh organic material would then either be exported as sinking particles or locally recycled, sustaining heterotrophic production of ammonium that would, in turn, be consumed by photoautotrophs in the euphotic layer but nitrified below. The resulting nitrite, eventually oxidized to nitrate, might remain evident at subsurface as observed in the nitrite anomaly of the young ring detected here. This hypothesis was supported by numerical simulations of the Massachusetts Institute of Technology General Circulation Model (*33*), which resolved Agulhas rings, their phytoplankton populations, and associated nutrient cycling (Fig. 6A). We tracked 12 Agulhas rings in the ocean model and characterized their near-surface biogeochemical cycles (Fig. 6B) (*34*). As the rings moved westward, storms enhanced surface heat loss, stimulating convection and the entrainment of nitrate. In the model simulations, proliferation of phytoplankton generated subsurface nitrite, which persisted because phytoplankton were light-limited at depth and because nitrification was suppressed by light at the surface (*35*). The associated blooms were dominated by large opportunistic phytoplankton and nitrate-metabolizing *Synechococcus* spp. analogs, whereas populations of *Prochlorococcus* spp. analogs dominated the quiescent periods (*34*). Each of the 12 simulated Agulhas rings exhibited this pattern in response to surface forcing by weather systems, and all rings maintained a persistent

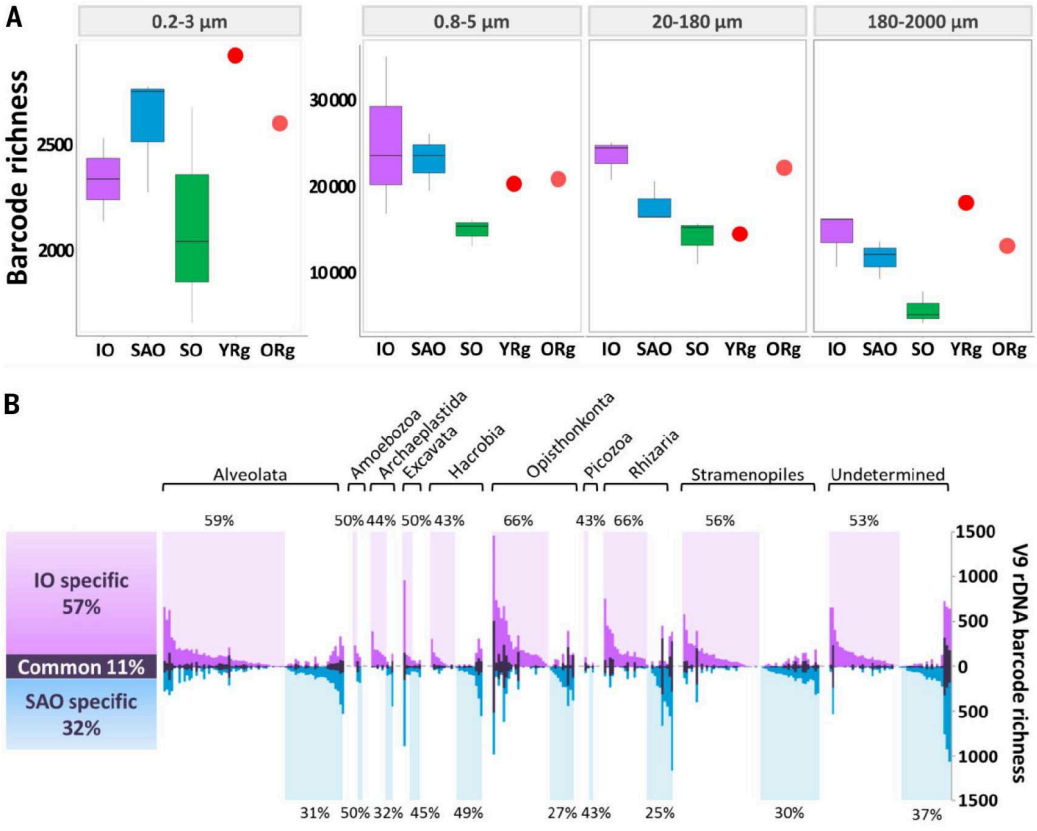


Fig. 3. Diversity of plankton populations specific to Indian and Atlantic Oceans. (**A**) Box plot of 16S (0.2 to 3 µm) and V9 rDNA barcodes richness (0.8- to 5-, 20- to 180-, and 180- to 2000-µm size fractions). Each box represents three sampling stations combined into Indian, South Atlantic, and Southern Ocean. Single Agulhas ring stations are represented as red (young ring) and orange (old ring) crosses. (**B**) Plankton sub-OTU richness filtering across the Agulhas choke point. Each vertical bar represents a single eukaryotic plankton OTU, each of which contains >10 distinct V9 rDNA barcodes (*14*). For each OTU are represented the number of distinct barcodes (sub-OTU richness) found exclusively in the South Atlantic Ocean (blue), exclusively in the Indian Ocean (pink), and in both South Atlantic Ocean and Indian Ocean (gray). OTUs are grouped by taxonomic annotation (indicated above the bar plot). For each taxonomic group, the percentage of OTUs with higher sub-OTU richness in the Indian Ocean (shaded in pink) or in the South Atlantic Ocean (shaded in blue) is indicated, respectively, at the top and bottom of the bar plot. A total of 1307 OTUs are presented, representing 98% of total V9 rDNA barcode abundance.

subsurface nitrite maximum in the region, as observed in TARA_068 and in other biogeochemical surveys (*36*).

The nitrite peak observed at TARA_068 in the young Agulhas ring was associated with a differential representation of nitrogen metabolism genes between the ring and the surrounding South Atlantic and Indian Oceans metagenomes derived from 0.2- to 3-μm size fractions (Fig. 7) (*37*). Agulhas ring overrepresented KEGG (Kyoto Encyclopedia of Genes and Genomes) orthologs (KOs) were involved in both nitrification and denitrification, likely representing the overlap between plankton assemblages involved in the conversion of nitrate to nitrite on the one hand and in denitrification of the accumulating nitrite on the other. Distinct KOs involved in successive denitrification steps were found to be encoded by similar plankton taxa. For instance, KO10945 and KO10946 (involved in ammonium nitrification) and KO00368 (subsequently involved in nitrite to nitrous oxide denitrification) appeared mostly encoded by Nitrosopumilaceae archaea. KO00264 and KO01674 (involved in ammonium assimilation) were mostly assigned to eukaryotic Mamiellales, whereas the opposite KO00367 and KO00366 (involved in dissimilatory nitrite reduction to ammonium), followed by KO01725 (involved in ammonium assimilation), were encoded by picocyanobacteria. In the specific case of the picocyanobacteria, metagenomic reads corresponding to *nirA* genes showed that the observed young Agulhas ring KO00366 (dissimilatory nitrite reduction) enrichment was mainly due to the overrepresentation of genes from *Prochlorococcus* (Fig. 8B). This enrichment was found to be associated with a concomitant shift in population structure from *Prochlorococcus* high-light II ecotypes (HLII, mostly lacking *nirA* genes) to codominance of high-light I (HLI) and low-light I (LLI) ecotypes. Indeed, among the several *Prochlorococcus* and *Synechococcus* ecotypes identified based on their genetic diversity and physiology (*38*, *39*), neutral marker (*petB*) (Fig. 8A) recruitments showed that dominant clades in the Indian Ocean upper mixed layer were *Prochlorococcus* HLII and *Synechococcus* clade II, as expected given the known (sub)tropical preference of these groups (*40*). Both clades nearly completely disappeared (less than 5%) in the mixed cold waters of the young ring and only began to increase again when the surface water warmed up along the South Atlantic Ocean transect. Conversely, young ring water was characterized by a large proportion of *Prochlorococcus* HLI and LLI and *Synechococcus* clade IV, two clades typical of temperate waters. Besides temperature, the *Prochlorococcus* community shift from HLII to HLI + LLI observed in the young ring was likely also driven by the nitrite anomaly. Indeed, whereas most *Synechococcus* strains isolated so far are able to

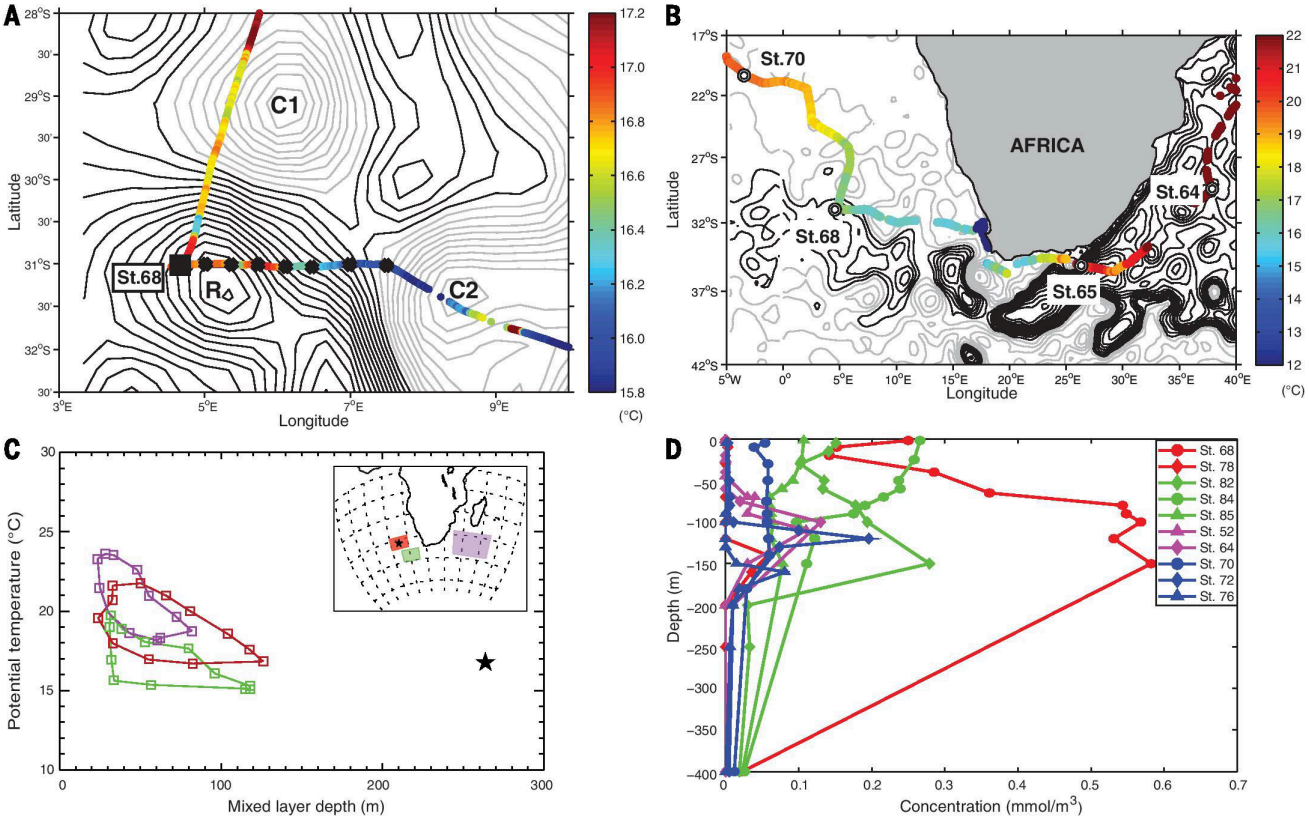


Fig. 4. Properties of the young Agulhas ring (TARA_068). (**A**) Daily sea surface height around young Agulhas ring station TARA_068 [absolute dynamic topography (ADT) from www.aviso.altimetry.fr]. R, C1, and C2, respectively, denote the centers of the Agulhas ring and two cyclonic eddies. The contour interval is 0.02 dyn/m. The ADT values are for 13 September 2010. Light gray isolines, ADT < 0.46 dyn/m. The crosses indicate the CTD stations, and the square symbol indicates the position of the biological station TARA_068. The biological station coincides with the westernmost CTD station. ADT is affected by interpolation errors, which is why CTD casts were performed at sea so as to have a fine-scale description of the feature before defining the position of the biological station (*23*). Superimposed are the continuous underway temperatures (°C) from the on-board thermosalinograph. (**B**) Same as (A) but at the regional scale. Round symbols correspond to biological sampling stations. The contour interval is 0.1 dyn/m. (**C**) Seasonal distribution of the median values of the mixed layer depths and temperatures at 10 m (from ARGO) provided by the IFREMER/LOS Mixed Layer Depth Climatology L2 database (www.ifremer.fr/cerweb/deboyer/mld) updated to 27 July 2011. The mixed layer is defined using a temperature criterion. The star symbol represents the young ring station TARA_068. (Inset) Geographic position of the areas used to select the mixed layer and temperature data. The mixed layer depth measured at TARA_068 is outside the 90th percentile of the distribution of mixed layer depths for the same month for both the subtropical (red and magenta) regions. The temperature matches the median for the same month and region of sampling. (**D**) Nitrite (NO_2) concentrations from CTD casts at different sampling sites (expressed in mmol/m^3).

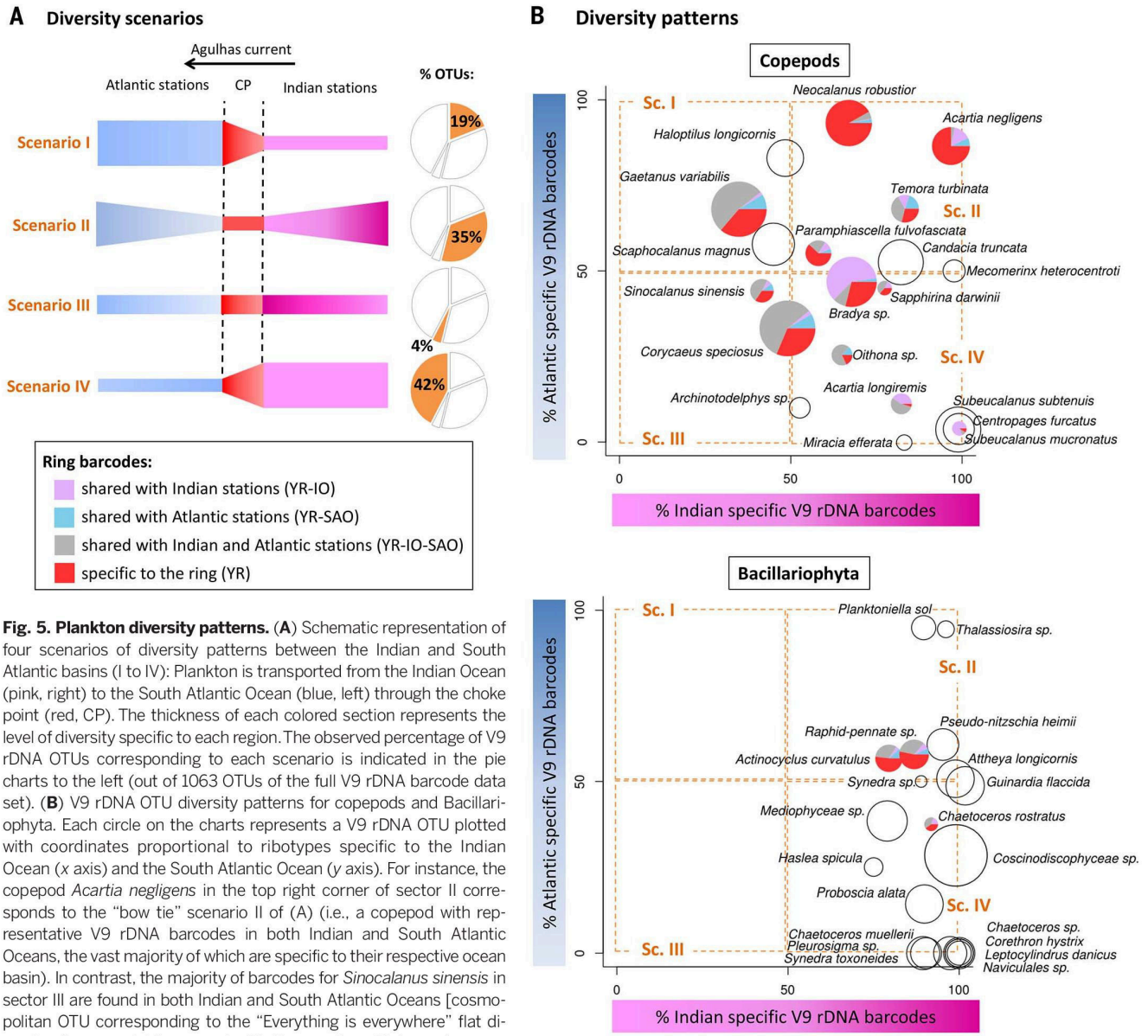


Fig. 5. Plankton diversity patterns. (**A**) Schematic representation of four scenarios of diversity patterns between the Indian and South Atlantic basins (I to IV): Plankton is transported from the Indian Ocean (pink, right) to the South Atlantic Ocean (blue, left) through the choke point (red, CP). The thickness of each colored section represents the level of diversity specific to each region. The observed percentage of V9 rDNA OTUs corresponding to each scenario is indicated in the pie charts to the left (out of 1063 OTUs of the full V9 rDNA barcode data set). (**B**) V9 rDNA OTU diversity patterns for copepods and Bacillariophyta. Each circle on the charts represents a V9 rDNA OTU plotted with coordinates proportional to ribotypes specific to the Indian Ocean (*x* axis) and the South Atlantic Ocean (*y* axis). For instance, the copepod *Acartia negligens* in the top right corner of sector II corresponds to the "bow tie" scenario II of (A) (i.e., a copepod with representative V9 rDNA barcodes in both Indian and South Atlantic Oceans, the vast majority of which are specific to their respective ocean basin). In contrast, the majority of barcodes for *Sinocalanus sinensis* in sector III are found in both Indian and South Atlantic Oceans [cosmopolitan OTU corresponding to the "Everything is everywhere" flat diversity diagram of (A), scenario III]. If more than 10 barcodes were found in the young Agulhas ring (TARA_068), their distribution is indicated in a pie chart (colors are coded in the legend inset); otherwise, the OTU is represented by an empty circle. Circle sizes are proportional to the number of considered barcodes for each OTU. The Bacillariophyta OTU defined as *Raphid pennate* sp. likely corresponds to the *Pseudo-nitzschia* cells observed by light microscopy.

use nitrate, nitrite, and ammonium, only the *Prochlorococcus* LLI and IV and some populations of HL clades, having acquired the *nirA* gene by lateral gene transfer, are able to assimilate nitrite. In the young ring, overrepresentation of cyanobacterial orthologs involved in nitrite reduction could thus have resulted from environmental pressure selecting LLI (87% of the *nirA* recruitments) and HL populations (13%) that possessed this ability. Because the capacity to assimilate nitrite in this latter ecotype reflects the availability of this nutrient in the environment (*41*), these in situ observations of picocyanobacteria indicated that the nitrogen cycle disturbance occurring in the young ring exerts community-wide selective pressure on Agulhas ring plankton.

Discussion

We found that whether or not the Agulhas choke point is considered a barrier to plankton dispersal depends on the taxonomic resolution at which the analysis is performed. At coarse taxonomic resolution, our observations of Indo-Atlantic continuous plankton structure—from viruses to fish larvae—suggested unlimited dispersal, consistent with previous reports (*5*, *42*). However, at finer resolution, our genetic data revealed that the Agulhas choke point strongly affects patterns of plankton genetic diversity. As anticipated in (*5*), the diversity filtering by Agulhas rings likely escaped detection using fossil records because of the limited taxonomic resolution afforded by fossil diatom morphology (*42*). The community-wide evidence presented here confirms observations on individual living species (*43*, *44*), suggesting that dispersal filters mitigate the panmictic ocean hypothesis for plankton above 20 μm.

The lower diversity we observed in the South Atlantic Ocean for micro- and mesoplankton (>20 μm) may be due to local abiotic/biotic pressure or to limitations in dispersal (*33*, *45*). Biogeography emerging from a model with only neutral drift (*46*) predicts

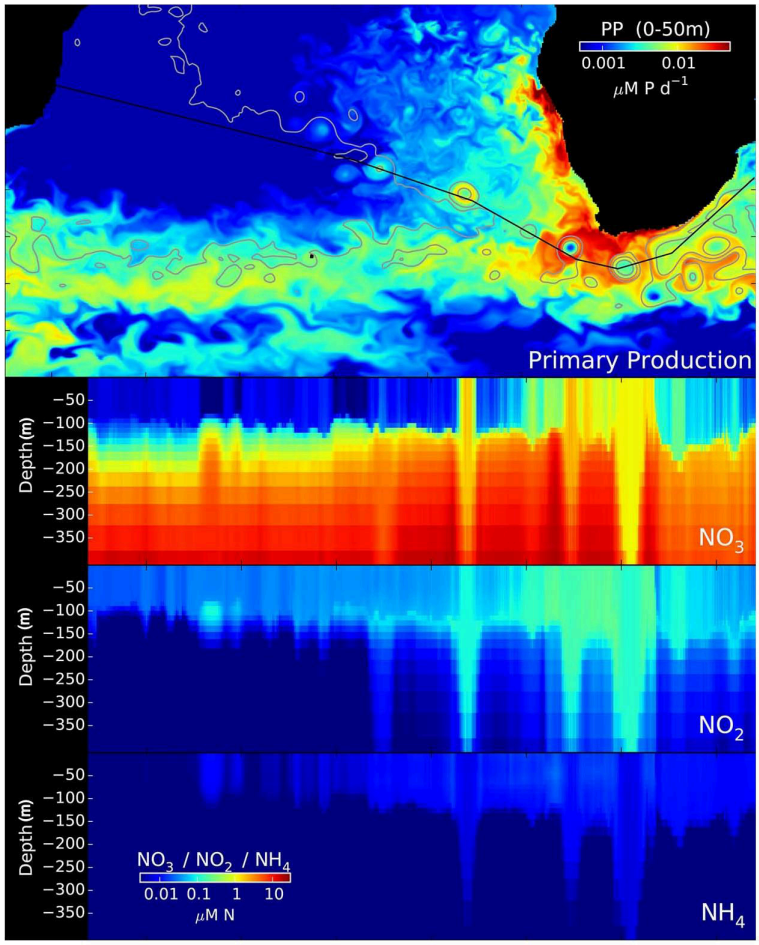


Fig. 6. Modeled nitrogen stocks along Agulhas ring track. (**Top**) Simulated primary production (PP) in the Agulhas system using the MIT-GCM model. The solid black line shows the average northwesterly path of 12 distinct virtual Agulhas rings tracked over the course of the simulation. Color scale for PP is given in the top right inset, with warmer colors indicating higher PP. (**Bottom**) Modeled profiles of NO_3, NO_2, and NH_4 along the Agulhas ring average track (*x* axis) presented in (A). The *y* axis is the depth (in meters) in the water column. The color scale is given in the bottom left inset, with warmer colors indicating higher concentrations of nitrogen compounds.

basin-to-basin genetic differences that are qualitatively consistent with our data. However, the increased proportion of *Prochlorococcus* HL populations carrying the *nirA* gene in the young Agulhas ring indicates that selection is at work in Agulhas rings. Based on our analysis of two Agulhas rings, we propose that environmental disturbances in Agulhas rings reshape their plankton diversity as they travel from the Indian Ocean to the South Atlantic Ocean. Such selective pressure may contribute to the South Atlantic Ocean plankton diversity shift relative to its upstream Indo-Pacific basin. Thus, environmental selection applied at a choke point in ocean circulation may constitute a barrier to dispersal (*47*, *48*). Furthermore, we show that taxonomic groups were not equally affected by the ring transport, both within and between phyla, with a noticeable effect of organism size. The differential effects due to organism size highlight the difficulty in generalizing ecological and evolutionary rules from limited sampling of species or functional types.

Considering the sensitivity of Agulhas leakage to climate change (*1*, *49*), better understanding of the plankton dynamics in Agulhas rings will be required if we are to understand and predict ecosystem resilience at the planetary scale. Considering the breadth of changes already observed in the 9-month-old Agulhas ring, it would be interesting to acquire samples from specific Agulhas rings tracked from early formation to dissipation. Finally, our data suggest that the abundance of Indian Ocean species in South Atlantic Ocean sedimentary records, used as proxies of Agulhas leakage intensity (*4*), may actually also depend on the physical and biological characteristics of the Agulhas rings.

Materials and methods

Sampling

The *Tara* Oceans sampling protocols schematized in Karsenti *et al.* (*9*) are described in Pesant *et al.* (*50*); specific methods for 0.8- to 5-, 20- to 180-, and 180- to 2000-μm size fractions in de Vargas *et al.* (*17*); for 0.2- to 3-μm size fractions in Sunagawa *et al.* (*51*); and for <0.2-μm size fraction in Brum *et al.* (*52*). Due to their fragility, 1.6-μm glass fiber filters initially used for prokaryote sampling were replaced by more resistant 3-μm polycarbonate filters from station TARA_066 onward. In the present text, both 0.2- to 1.6-μm and 0.2- to 3-μm prokaryote size fractions are simply referred to as 0.2 to 3 μm.

Data acquisition

A range of analytical methods covering different levels of taxonomic resolution (pigments, flow cytometry, optical microscopy, marker gene barcodes, and metagenomics) were used to describe the planktonic composition at each sampled station. Viruses from the <0.2 μm size fraction were studied by epifluorescence microscopy, by quantitative transmission electron microscopy, and by sequencing DNA as described in Brum *et al.* (*52*). Flow cytometry was used to discriminate high-DNA-content bacteria (HNA), low-DNA-content bacteria (LNA), *Prochlorococcus* and *Synechococcus* picocyanobacteria, and two different groups (based on their size) of photosynthetic picoeukaryotes, as described previously (*53*). Pigment concentrations measured by high-performance liquid chromatography (HPLC) were used to estimate the dominant classes of phytoplankton using the CHEMTAX procedure (*54*). Tintinnids, diatoms, and dinoflagellates were identified and counted by light microscopy from the 20- to 180-μm lugol or formaldehyde fixed-size fraction. Zooplankton enumeration was performed on formol fixed samples using the ZOOSCAN semi-automated classification of digital images (*55*). Sequencing, clustering, and annotation of 18*S*-V9 rDNA barcodes are described in de Vargas *et al.* (*17*). Metagenome sequencing, assembly, and annotation are described in Sunagawa *et al.* (*51*). NCLDV taxonomic assignations in the 0.2- to 3-μm samples were carried out using 18 lineage-specific markers as described in Hingamp *et al.* (*56*). Virome sequencing and annotation are described in Brum *et al.* (*52*). Samples and their associated contextual data are described at PANGAEA (*57*–*59*).

Data analysis

Origin of sampled Agulhas rings

Using visual and automated approaches, the origins of the TARA_068 and TARA_078 stations were traced back from the daily altimetric data (Fig. 1) (*21*). The automated approach used either the Lagrangian tracing of numerical particles initialized in the center of a given structure and transported by the geostrophic velocity field calculated from sea surface height gradients, or the connection in space and time of adjacent extreme values in sea level anomaly maps.

V9 rDNA barcodes

To normalize for differences in sequencing effort, V9 rDNA barcode libraries were resampled 50 times for the number of reads corresponding to the smallest library in each size fraction: 0.8 to 5 μm, 776,358 reads; 20 to 180 μm, 1,170,592 reads; and 180 to 2000 μm, 767,940 reads. V9 rDNA barcode counts were then converted to the average number

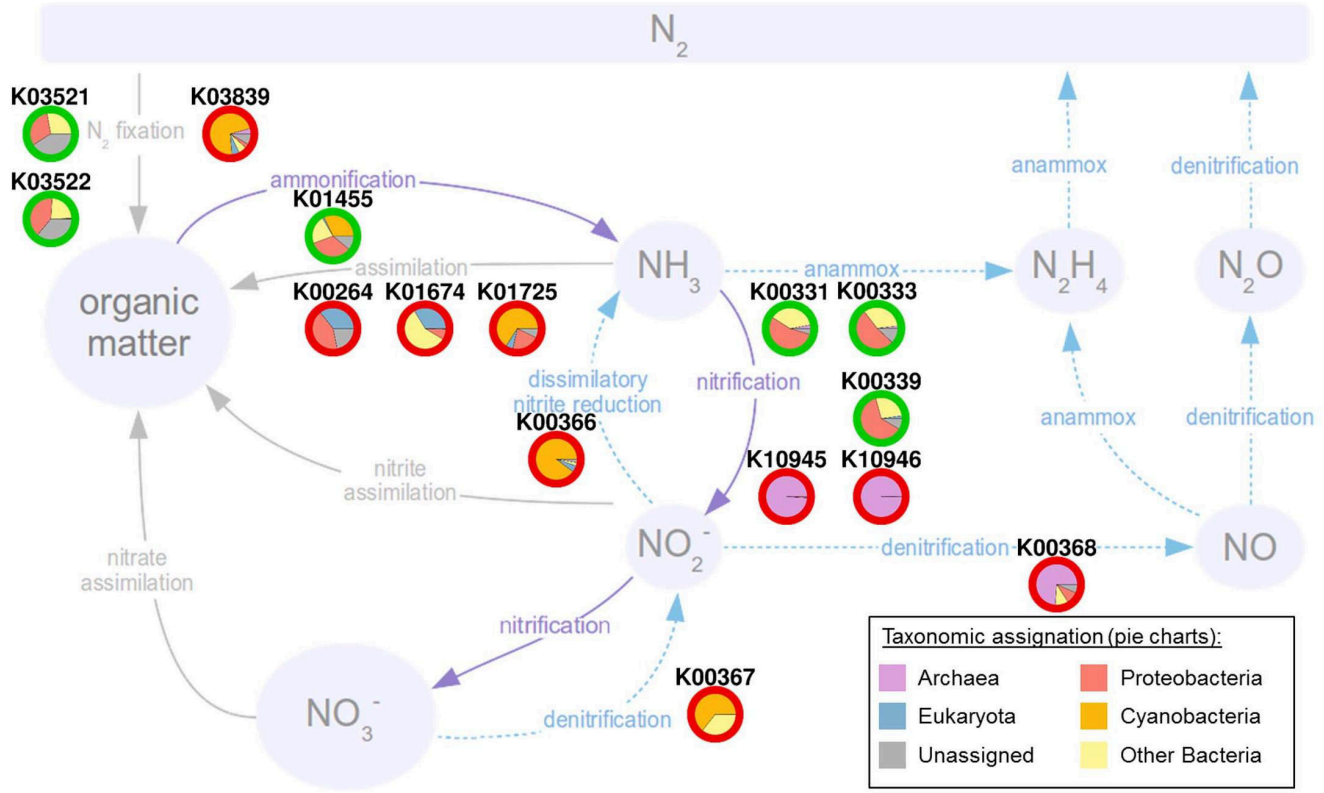


Fig. 7. Nitrite anomaly in the young Agulhas ring is accompanied by shifts in nitrogen pathway–related genes. Metagenomic over- and underrepresented nitrogen pathway genes in young Agulhas ring. Over- (red circles) and under- (green circles) represented metagenome functional annotations (KEGG Orthologs, KO#) involved in the nitrogen pathway in the young ring compared to Indian and South Atlantic Oceans reference stations, at surface and deep chlorophyll maximum depth. Pie charts inside circles represent the taxonomic distribution for each ortholog.

of times seen in the 50 resampling events, and barcodes with less than 10 reads were removed as potential sequencing artifacts. We used down-sampled barcode richness (number of distinct V9 rDNA barcodes) as a diversity descriptor because using V9 rDNA barcode abundances to compare plankton assemblages would likely be biased due to (i) technical limitations described in de Vargas *et al.* (*17*) and (ii) seasonality effects induced by the timing of samplings (table S1). Barcode richness was well correlated with Shannon and Simpson indexes (0.94 and 0.78, respectively). The shared barcode richness between each pair of samples (*14*) was estimated by counting, for the three larger size fractions (0.8 to 5, 20 to 180, and 180 to 2000 μm), the proportion of V9 rDNA barcodes 100% identical over their whole length. V9 rDNA barcodes were clustered into OTUs by swarm clustering as described by de Vargas *et al.* (*17*). The sub-OTU richness comparison between two samples s1 and s2 (*14*) produces three values: the number of V9 rDNA barcodes in common, the number of V9 rDNA barcodes unique to s1, and the number of V9 rDNA barcodes unique to s2. These numbers can be represented directly as bar graphs (Fig. 3B) or as dot plots of specific V9 rDNA barcode richness (Fig. 5).

Metagenomic analysis

Similarity was estimated using whole shotgun metagenomes for all four available size fractions (0.2 to 3, 0.8 to 5, 20 to 180, and 180 to 2000 μm). Because pairwise comparisons of all raw metagenome reads are intractable given the present data volume, we used a heuristic in which two metagenomic 100–base pair (bp) reads were considered similar if at least two nonoverlapping 33-bp subsequences were strictly identical (Compareads method) (*60*). For prokaryotic fractions (0.2 to 3 μm), taxonomic abundance was estimated using the number of $16S_{mi}$tags (*51*). The functional annotation, taxonomic assignation, and gene abundance estimation of the panoceanic Ocean Microbial Reference Gene Catalog (OM-RGC) (243 samples, including all those analyzed here) generated from *Tara* Oceans 0.2- to 3-μm metagenomic reads are described in Sunagawa *et al.* (*51*). Gene abundances were computed for the set of genes annotated to the nitrogen metabolism KO (*61*) group by counting the number of reads from each sample that mapped to each KO-associated gene. Abundances were normalized as reads per kilobase per million mapped reads (RPKM). Gene abundances were then aggregated (summed) for each KO group. To compare abundances between the young ring (TARA_068) and other stations, a *t* test was used. KOs with a *P* value <0.05 and a total abundance (over all stations) >10 were considered as significant (*37*). *Prochlorococcus* and *Synechococcus* community composition was analyzed in the 0.2- to 3-μm size fraction at the clade level by recruiting reads targeting the high-resolution marker gene *petB*, coding for cytochrome b_6 (*62*). The *petB* reads were first extracted from metagenomes using Basic Local Alignment Search Tool (BLASTx+) against the *petB* sequences of *Synechococcus* sp. WH8102 and *Prochlorococcus marinus* MED4. These reads were subsequently aligned against a reference data set of 270 *petB* sequences using BLASTn (with parameters set at -G 8 -E 6 -r 5 -q -4 -W 8 -e 1 -F "m L" -U T). *petB* reads exhibiting >80% identity over >90% of sequence length were then taxonomically assigned to the clade of the best BLAST hit. Read counts per clade were normalized based on the sequencing effort for each metagenomic sample. A similar approach was used with *nirA* (KO 00366) and *narB* genes (KO 00367), which were highlighted in the nitrogen-related KO analysis (Fig. 7). Phylogenetic assignment was realized at the highest possible taxonomic level using a reference data set constituted of sequences retrieved from Cyanorak v2 (www.sb-roscoff.fr/cyanorak/) and Global Ocean Sampling (*41*, *63*) databases.

Nitrogen cycle modeling

Numerical simulations of global ocean circulation were based on the Massachusetts Institute of Technology General Circulation Model (MIT-GCM) (*64*), incorporating biogeochemical and ecological components (*65*, *66*). It resolved mesoscale

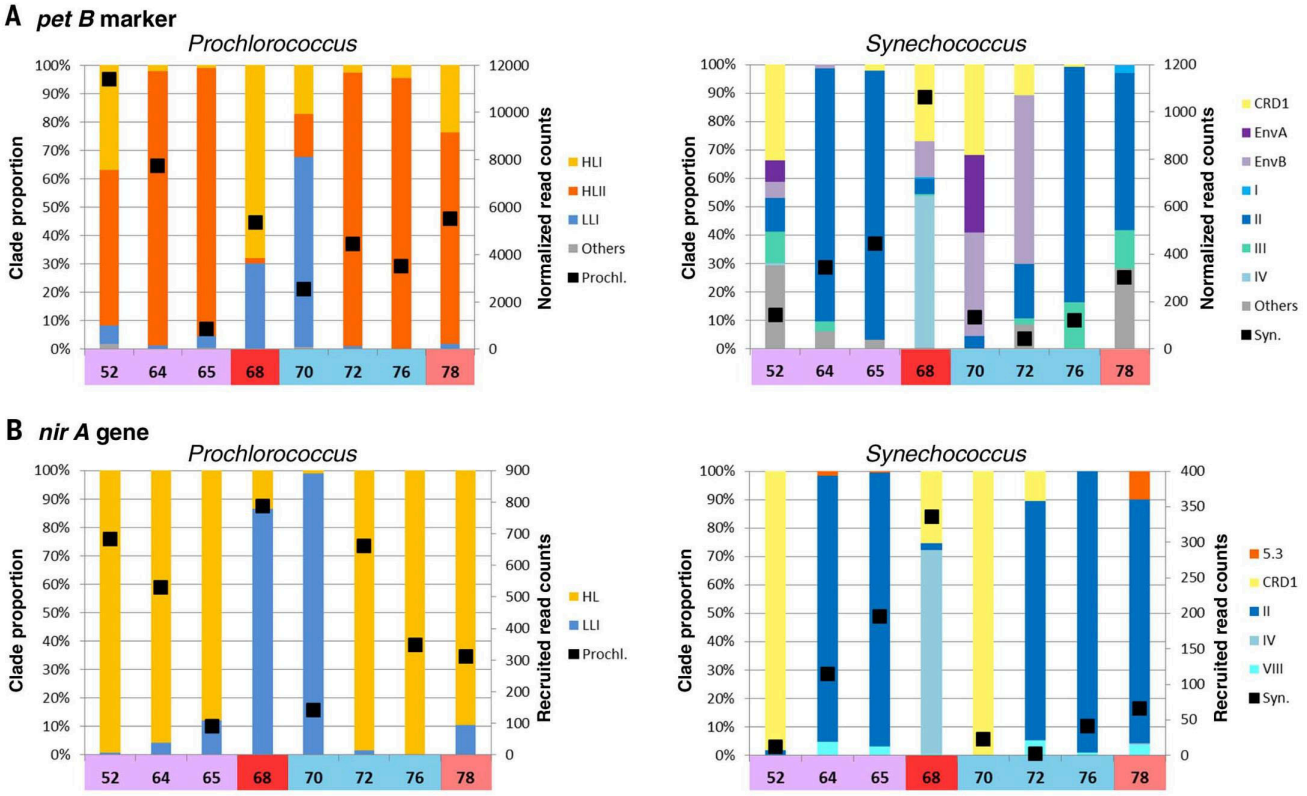


Fig. 8. Picocyanobacterial clade shift in the young Agulhas ring. (**A**) Relative abundance of *Prochlorococcus* and *Synechococcus* clades, estimated by *petB* read recruitments from 0.2- to 3-μm metagenomes. Solid squares correspond to read counts normalized based on the sequencing effort (right axis). (**B**) Relative abundance of *nirA* gene from *Prochlorococcus* and *Synechococcus* clades estimated by number of reads recruited from 0.2- to 3-μm metagenomes. The bar colors correspond to cyanobacterial clades indicated in the inset legends for each panel. Solid squares correspond to the number of reads recruited (right axis). Data are shown for stations TARA_052 to TARA_078 only, because too few cyanobacteria were found in Southern Ocean stations TARA_082, TARA_084, and TARA_085.

features in the tropics and was eddy-permitting in subpolar regions. The physical configurations were integrated from 1992 to 1999 and constrained to be consistent with observed hydrography and altimetry (*67*). Three inorganic fixed nitrogen pools were resolved—nitrate, nitrite, and ammonium—as well as particulate and dissolved detrital organic nitrogen. Phytoplankton types were able to use some or all of the fixed nitrogen pools. Aerobic respiration and remineralization by heterotrophic microbes was parameterized as a simple sequence of transformations from detrital organic nitrogen, to ammonium, then nitrification to nitrite and nitrate. In accordance with empirical evidence (*35*), nitrification was assumed to be inhibited by light. Nitrification is described in the model by simple first-order kinetics, with rates tuned to qualitatively capture the patterns of nitrogen species in the Atlantic (*66*).

Continuous spectral analysis

A continuous flow-through system equipped with a high-spectral-resolution spectrophotometer (AC-S, WET Labs, Inc.) was used for data collection during the *Tara* Oceans expedition, as described previously (*68*). Phytoplankton pigment concentrations, estimates of phytoplankton size γ, total chlorophyll a concentration, and particulate organic carbon (POC) are derived from the absorption and attenuation spectra (*69*) for the 1-km^2-binned *Tara* Oceans data set available at PANGAEA (http://doi.pangaea.de/10.1594/PANGAEA.836318).

REFERENCES AND NOTES

1. A. Biastoch, C. W. Böning, J. R. E. Lutjeharms, Agulhas leakage dynamics affects decadal variability in Atlantic overturning circulation. *Nature* **456**, 489–492http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=19037313&dopt=Abstract (2008). doi: 10.1038/nature07426
2. A. Biastoch, C. W. Böning, F. U. Schwarzkopf, J. R. E. Lutjeharms, Increase in Agulhas leakage due to poleward shift of Southern Hemisphere westerlies. *Nature* **462**, 495–498 http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=19940923&dopt=Abstract (2009). doi: 10.1038/nature08519
3. L. M. Beal *et al*., On the role of the Agulhas system in ocean circulation and climate. *Nature* **472**, 429–436 (2011). doi: 10.1038/nature09983; pmid: 21525925
4. F. J. C. Peeters *et al*., Vigorous exchange between the Indian and Atlantic oceans at the end of the past five glacial periods. *Nature* **430**, 661–665 (2004). doi: 10.1038/nature02785; pmid: 15295596
5. P. Cermeño, P. G. Falkowski, Controls on diatom biogeography in the ocean. *Science* **325**, 1539–1541 (2009). pmid: 19762642
6. A. L. Gordon, Oceanography: The brawniest retroflection. *Nature* **421**, 904–905 (2003). doi: 10.1038/421904a; pmid: 12606984
7. H. M. van Aken *et al*., Observations of a young Agulhas ring, Astrid, during MARE in March 20. *Deep Sea Res. Part II Top. Stud. Oceanogr.* **50**, 167–195 (2003). doi: 10.1016/S0967-0645(02)00383-1
8. J. R. Bernhardt, H. M. Leslie, Resilience to climate change in coastal marine ecosystems. *Annu. Rev. Mar. Sci.* **5**, 371–392 (2013). doi: 10.1146/annurev-marine-121211-172411; pmid: 22809195
9. E. Karsenti *et al*., A holistic approach to marine eco-systems biology. *PLOS Biol.* **9**, e1001177 (2011). doi: 10.1371/journal.pbio.1001177; pmid: 22028628
10. Companion Web site, tables W2 and W3; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#TablesW
11. A. H. Orsi, T. Whitworth III, W. D. Nowlin Jr., On the meridional extent and fronts of the Antarctic Circumpolar Current. *Deep-Sea Res.* **42**, 641–673 (1995). doi: 10.1016/0967-0637(95)00021-W
12. Companion Web site, tables W4 to W12; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#TablesW
13. Companion Web site, figure W1; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW1
14. Companion Web site, figure W2; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW2
15. Companion Web site, tables W13 and W14; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#TablesW
16. Companion Web site, figure W3; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW3
17. C. de Vargas *et al*., Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).
18. B. W. Bowen, L. A. Rocha, R. J. Toonen, S. A. KarlToBo Laboratory, The origins of tropical marine biodiversity. *Trends Ecol. Evol.* **28**, 359–366 (2013). doi: 10.1016/j.tree.2013.01.018; pmid: 23453048
19. R. L. Cunha *et al*., Ancient divergence in the trans-oceanic deep-sea shark *Centroscymnus crepidater*. *PLOS ONE* **7**, e49196 (2012). doi: 10.1371/journal.pone.0049196; pmid: 23145122
20. C. Schmid *et al*., Early evolution of an Agulhas Ring. *Deep Sea Res. Part II Top. Stud. Oceanogr.* **50**, 141–166 (2003). doi: 10.1016/S0967-0645(02)00382-X

21. Companion Web site, figure W4; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW4
22. Companion Web site, figure W5; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW5
23. Companion Web site, figure W6; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW6
24. Companion Web site, figure W7; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW7
25. L. Gordon, J. R. Lutjeharms, M. L. Gründlingh, Stratification and circulation at the Agulhas Retroflection. *Deep-Sea Res. A, Oceanogr. Res. Pap.* **34**, 565–599 (1987). doi: 10.1016/0198-0149(87)90006-9
26. V. Faure, M. Arhan, S. Speich, S. Gladyshev, Heat budget of the surface mixed layer south of Africa. *Ocean Dyn.* **61**, 1441–1458 (2011). doi: 10.1007/s10236-011-0444-1
27. Companion Web site, figure W8; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW8
28. Companion Web site, figure W9; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW9
29. Companion Web site, text W1; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#TextW1
30. Companion Web site, figure W10; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW10
31. Companion Web site, figure W11; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW11
32. Companion Web site, figure W12; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW12
33. S. Clayton, S. Dutkiewicz, O. Jahn, M. J. Follows, Dispersal, eddies, and the diversity of marine phytoplankton. *Limnol. Oceanogr. Fluids Environ.* **3**, 182–197 (2013). doi: 10.1215/21573689-2373515
34. Companion Web site, figure W13; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW13
35. R. J. Olson, Differential photoinhibition of marine nitrifying bacteria: A possible mechanism for the formation of the primary nitrite maximum. *J. Mar. Res.* **39**, 227–238 (1981).
36. S. Levitus *et al.*, The World Ocean Database. *Data Sci. J.* **12**, WDS229–WDS234 (2013). doi: 10.2481/dsj.WDS-041
37. Companion Web site, table W15; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#TablesW
38. D. J. Scanlan *et al.*, Ecological genomics of marine picocyanobacteria. *Microbiol. Mol. Biol. Rev.* **73**, 249–299 (2009). doi: 10.1128/MMBR.00035-08; pmid: 19487728
39. Z. I. Johnson *et al.*, Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. *Science* **311**, 1737–1740 (2006). doi: 10.1126/science.1118052; pmid: 16556835
40. K. Zwirglmaier *et al.*, Global phylogeography of marine Synechococcus and Prochlorococcus reveals a distinct partitioning of lineages among oceanic biomes. *Environ. Microbiol.* **10**, 147–161 (2008). pmid: 17900271
41. A. C. Martiny, S. Kathuria, P. M. Berube, Widespread metabolic potential for nitrite and nitrate assimilation among *Prochlorococcus* ecotypes. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 10787–10792 (2009). doi: 10.1073/pnas.0902532106; pmid: 19549842
42. C. Hubert *et al.*, A constant flux of diverse thermophilic bacteria into the cold Arctic seabed. *Science* **325**, 1541–1544 (2009). doi: 10.1126/science.1174012; pmid: 19762643
43. C. K. C. Churchill, Á. Valdés, D. Ó. Foighil, Afro-Eurasia and the Americas present barriers to gene flow for the cosmopolitan neustonic nudibranch *Glaucus atlanticus*. *Mar. Biol.* **161**, 899–910 (2014). doi: 10.1007/s00227-014-2389-7
44. N. Selje, M. Simon, T. Brinkhoff, A newly discovered *Roseobacter* cluster in temperate and polar oceans. *Nature* **427**, 445–448 (2004). doi: 10.1038/nature02272; pmid: 14749832
45. G. Casteleyn *et al.*, Limits to gene flow in a cosmopolitan marine planktonic diatom. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 12952–12957 (2010). doi: 10.1073/pnas.1001380107; pmid: 20615950
46. F. L. Hellweger, E. van Sebille, N. D. Fredrick, Biogeographic patterns in ocean microbes emerge in a neutral agent-based model. *Science* **345**, 1346–1349 (2014). doi: 10.1126/science.1254421; pmid: 25214628
47. D. H. Janzen, Why mountain passes are higher in the tropics. *Am. Nat.* **101**, 233–249 (1967). doi: 10.1086/282487
48. G. Wang, M. E. Dillon, Recent geographic convergence in diurnal and annual temperature cycling flattens global thermal profiles. *Nature Climate Change* **4**, 988–992 (2014). doi: 10.1038/nclimate2378
49. C. Backeberg, P. Penven, M. Rouault, Impact of intensified Indian Ocean winds on mesoscale variability in the Agulhas system. *Nature Clim. Change* **2**, 608–612 (2012). doi: 10.1038/nclimate1587
50. S. Pesant *et al.*, Open science resources for the discovery and analysis of *Tara* Oceans data. http://biorxiv.org/content/early/2015/05/08/019117 (2015).
51. S. Sunagawa *et al.*, Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
52. J. R. Brum *et al.*, Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498 (2015).
53. J. M. Gasol, P. A. Del Giorgio, Using flow cytometry for counting natural planktonic bacteria and understanding the structure of planktonic bacterial communities. *Sci. Mar.* **64**, 197–224 (2000).
54. M. Mackey, D. Mackey, H. Higgins, S. Wright, CHEMTAX - a program for estimating class abundances from chemical markers: Application to HPLC measurements of phytoplankton. *Mar. Ecol. Prog. Ser.* **144**, 265–283 (1996). doi: 10.3354/meps144265
55. G. Gorsky *et al.*, Digital zooplankton image analysis using the ZooScan integrated system. *J. Plankton Res.* **32**, 285–303 (2010). doi: 10.1093/plankt/fbp124
56. P. Hingamp *et al.*, Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J.* **7**, 1678–1695 (2013). doi: 10.1038/ismej.2013.59; pmid: 23575371
57. Tara Oceans Consortium Coordinators; Tara Oceans Expedition, Participants (2014): Registry of selected samples from the Tara Oceans Expedition (2009-2013). doi: 10.1594/PANGAEA.840721
58. S. Chaffron, L. Guidi, F. D'Ovidio, S. Speich, S. Audic, S. De Monte, D. Iudicone, M. Picheral, S. Pesant; Tara Oceans Consortium Coordinators, Tara Oceans Expedition, Participants (2014): Contextual environmental data of selected samples from the Tara Oceans Expedition (2009-2013). doi: 10.1594/PANGAEA.840718
59. S. Chaffron, F. D'Ovidio, S. Sunagawa, S. G. Acinas, L. P. Coelho, S. De Monte, G. Salazar, S. Pesant; Tara Oceans Consortium Coordinators, Tara Oceans Expedition, Participants (2014): Contextual biodiversity data of selected samples from the Tara Oceans Expedition (2009-2013). doi: 10.1594/PANGAEA.840698
60. N. Maillet, C. Lemaitre, R. Chikhi, D. Lavenier, P. Peterlongo, Compareads: Comparing huge metagenomic experiments. *BMC Bioinformatics* **13** (suppl. 19), S10 (2012). doi: 10.1186/1471-2105-13-S19-S10; pmid: 23282463
61. M. Kanehisa *et al.*, KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36** (Database), D480–D484 (2008). doi: 10.1093/nar/gkm882; pmid: 18077471
62. S. Mazard, M. Ostrowski, F. Partensky, D. J. Scanlan, Multi-locus sequence analysis, taxonomic resolution and biogeography of marine Synechococcus. *Environ. Microbiol.* **14**, 372–386 (2012). doi: 10.1111/j.1462-2920.2011.02514.x; pmid: 21651684
63. D. B. Rusch *et al.*, The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLOS Biol.* **5**, e77 (2007). doi: 10.1371/journal.pbio.0050077; pmid: 17355176
64. J. Marshall, A. Adcroft, C. Hill, L. Perelman, C. Heisey, A finite-volume, incompressible Navier Stokes model for studies of the ocean on parallel computers. *J. Geophys. Res.* **102** (C3), 5753–5766 (1997). doi: 10.1029/96JC02775
65. M. J. Follows, S. Dutkiewicz, S. Grant, S. W. Chisholm, Emergent biogeography of microbial communities in a model ocean. *Science* **315**, 1843–1846 (2007). doi: 10.1126/science.1138544; pmid: 17395828
66. S. Dutkiewicz, M. J. Follows, J. G. Bragg, Modeling the coupling of ocean ecology and biogeochemistry. *Global Biogeochem. Cycles* **23**, GB4017 (2009). doi: 10.1029/2008GB003405
67. D. Menemenlis *et al.*, ECCO2: High resolution global ocean and sea ice data synthesis. *Mercator Ocean Quarterly Newsletter* **31**, 13–21 (2008).
68. A. Chase *et al.*, Decomposition of in situ particulate absorption spectra. *Methods in Oceanography.* **7**, 110–124 (2013). doi: 10.1016/j.mio.2014.02.002
69. E. Boss *et al.*, The characteristics of particulate absorption, scattering and attenuation coefficients in the surface ocean; Contribution of the Tara Oceans expedition. *Methods in Oceanography.* **7**, 52–62 (2013). doi: 10.1016/j.mio.2013.11.002
70. K. R. Ridgway, J. R. Dunn, J. L. Wilkin, Ocean interpolation by four-dimensional least squares - Application to the waters around Australia. *J. Atmos. Ocean. Technol.* **19**, 1357–1375 (2002). doi: 10.1175/1520-0426(2002)019<1357:OIBFDW>2.0.CO;2

ACKNOWLEDGMENTS

We appreciate the commitment of the following people and sponsors: 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; the French government Investissements d'Avenir programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), MEMO LIFE (ANR-10-LABX-54), PSL* Research University (ANR-11-IDEX-0001-02), and ANR (projects POSEIDON/ANR-09-BLAN- 0348, PHYTBACK/ANR-2010-1709-01, PROMETHEUS/ANR-09-PCS-GENM-217, SAMOSA/ANR-13-ADAP-0010, and TARAGIRUS/ANR-09-PCS-GENM-218); European Union FP7 (MicroB3/No.287589, IHMS/HEALTH-F4-2010-261376, and MaCuMBA/No.311975); ERC Advanced Grant Award to C.B. (Diatomite: 294823); Gordon and Betty Moore Foundation grant (no. 3790) to M.B.S.; Spanish Ministry of Science and Innovation grant CGL2011-26848/BOS MicroOcean PANGENOMICS to S.G.A.; TANIT (CONES 2010-0036) from the Agència de Gestió d´Ajusts Universitaris i Reserca to S.G.A.; JSPS KAKENHI grant no. 26430184 to H.O.; NASA Ocean Biology and Biogeochemistry program (NNX11AQ14G and NNX09AU43G) to E.B.; The Italian Research for the Sea (Flagship Project RITMARE) to D.I.; and FWO, BIO5, and Biosphere 2 to M.B.S. We also appreciate the support and commitment of Agnès b. and Etienne Bourgois, the Veolia Environment Foundation, Region Bretagne, Lorient Agglomeration, World Courier, Illumina, the EDF Foundation, FRB, the Prince Albert II de Monaco Foundation, and the *Tara* schooner and its captains and crew. We thank MERCATOR-CORIOLIS and ACRI-ST for providing daily satellite data during the expedition. We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries that graciously granted sampling permissions. *Tara* Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org). We also acknowledge excellent assistance from the European Bioinformatics Institute (EBI), in particular G. Cochrane and P. ten Hoopen, as well as the EMBL Advanced Light Microscopy Facility (ALMF), in particular R. Pepperkok. We thank Y. Timsit for stimulating scientific discussions and critical help during writing of the manuscript. The altimeter products were produced by Ssalto/Duacs and CLS, with support from CNES. The authors further declare that all data reported herein are fully and freely available from the date of publication, with no restrictions, and that all of the samples, analyses, publications, and ownership of data are free from legal entanglement or restriction of any sort by the various nations whose waters the *Tara* Oceans expedition sampled. Data described herein are available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas, at EBI under the project identifiers PRJEB402 and PRJEB7988, and at PANGAEA (*57*–*59*). The data release policy regarding future public release of *Tara* Oceans data is described in Pesant *et al.* (*50*). All authors approved the final manuscript. This article is contribution number 21 of *Tara* Oceans. The supplementary materials contain additional data. See also http://doi.pangaea.de/10.1594/PANGAEA.840721; http://doi.pangaea.de/10.1594/PANGAEA.840718; and http://doi.pangaea.de/10.1594/PANGAEA.840698

***Tara* Oceans Coordinators**

Silvia G. Acinas,[1] Peer Bork,[2] Emmanuel Boss,[3] Chris Bowler,[4] Colomban de Vargas,[5,6] Michael Follows,[7] Gabriel Gorsky,[8,9] Nigel Grimsley,[10,11] Pascal Hingamp,[12] Daniele Iudicone,[13] Olivier Jaillon,[14,15,16] Stefanie Kandels-Lewis,[2,17] Lee Karp-Boss,[3] Eric Karsenti,[4,17] Uros Krzic,[18] Fabrice Not,[5,6] Hiroyuki Ogata,[19] Stephane Pesant,[20,21] Jeroen Raes,[22,23,24] Emmanuel G. Reynaud,[25] Christian Sardet,[26,27] Mike Sieracki,[28] Sabrina Speich,[29,30] Lars Stemmann,[8] Matthew B. Sullivan,[31] Shinichi Sunagawa,[2] Didier Velayoudon,[32] Jean Weissenbach,[14,15,16] Patrick Wincker[14,15,16]

[1]Department of Marine Biology and Oceanography, Institute of Marine Science (ICM)-CSIC, Pg. Marítim de la Barceloneta, 37-49, Barcelona E08003, Spain. [2]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [3]School of Marine Sciences, University of Maine, Orono, Maine, USA. [4]Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), and Inserm U1024, and CNRS UMR 8197, F-75005 Paris, France. [5]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [6]Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [7]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [8]CNRS, UMR 7093, LOV, Observatoire Océanologique, F-06230 Villefranche-sur-Mer, France. [9]Sorbonne Universités, UPMC Univ Paris 06, UMR 7093, LOV, Observatoire Océanologique, F-06230 Villefranche-sur-Mer, France. [10]CNRS UMR 7232, BIOM, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [11]Sorbonne Universités

Paris 06, OOB UPMC, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [12]Aix Marseille Université CNRS IGS UMR 7256, 13288 Marseille, France. [13]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [14]CEA, Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France. [15]CNRS, UMR 8030, CP5706, Evry, France. [16]Université d'Evry, UMR 8030, CP5706, Evry, France. [17]Directors' Research, European Molecular Biology Laboratory, Heidelberg, Germany. [18]Cell Biology and Biophysics, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [19]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan. [20]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany. [21]MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany. [22]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [23]Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. [24]Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [25]Earth Institute, University College Dublin, Dublin, Ireland. [26]CNRS, UMR 7009 Biodev, Observatoire Océanologique, F-06230 Villefranche-sur-Mer, France. [27]Sorbonne Universités, UPMC Univ Paris 06, UMR 7009 Biodev, F-06230 Observatoire Océanologique, Villefranche-sur-Mer, France. [28]Bigelow Laboratory for Ocean Sciences, East Boothbay, USA. [29]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond 75231 Paris Cedex 05, France. [30]Laboratoire de Physique des Océans, UBO-IUEM, Place Copernic, 29820 Plouzané, France. [31]Department of Ecology and Evolutionary Biology, University of Arizona, 1007 East Lowell Street, Tucson, AZ 85721, USA. [32]DVIP Consulting, Sèvres, France.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/1261447/suppl/DC1
Table S1

18 September 2014; accepted 23 February 2015
10.1126/science.1261447

Résumé (français) :

Les diatomées sont des micro-algues unicellulaires, qui jouent un rôle primordial dans l'eco-système marin. En effet, elles sont responsables de 20% de l'activité photosynthétique sur Terre, et sont à la base de la chaîne alimentaire marine, toujours plus menacée par le changement climatique.

Les diatomées établissent diverses interactions microbiennes avec des organismes issus de l'ensemble de l'arbre du vivant, à travers des méchanismes complexes tels que la symbiose, le parasitisme ou la compétition. L'objectif de ma thèse a été de comprendre comment ces interactions structurent la communauté du plancton, à grande échelle spatiale. Pour ce faire, j'ai développé de nouvelles approches basées sur le jeu de données inédit de *Tara* Océans, une expédition mondiale qui a exploré la diversité et les fonctions des microbes marins, en récoltant plus de 40.000 échantillons à travers 210 sites autour du monde.

Grâce à l'analyse de réseaux de co-occurrence microbiens, je montre d'une part que les diatomées agissent comme des « ségrégateurs répulsifs » à l'échelle globale, en particulier envers les organismes potentiellement dangereux tels que les prédateurs et les parasites, et d'autre part que la co-occurrence des espèces ne s'explique qu'en minorité par les facteurs environnementaux. Grâce à la richesse des données *Tara* Océans, j'ai par ailleurs permis la charactérisation d'une interaction biotique impliquant une diatomée et un cilié hétérotrophe à l'échelle de l'eco-système, illustrant de surcroît le succès des approches dirigées par les données. Dans l'ensemble, ma thèse contribue à notre compréhension des interactions biotiques impliquant les diatomées, de l'échelle globale à la cellule unique.

Title:
Diatom interactions in the open ocean: from the global patterns to the single cell.

Abstract :
Diatoms are unicellular photosynthetic microeukaryotes that play a critical role in the functioning of marine ecosystems. They are responsible for 20% of global photosynthesis on Earth and lie at the base of marine food webs, ever more threatened by climate change.

Diatoms establish microbial interactions with numerous organisms across the whole tree of life, through complex mechanisms including symbiosis, parasitism and competition. The goal

of my thesis was to understand how those biotic interactions structure the planktonic community at large spatial scales, by using new approaches based on the unprecedented *Tara* Oceans dataset, a unique and worldwide circumnavigation that collected over 40.000 samples across 210 sites to explore the diversity and functions of marine microbes.

Through the analysis of microbial association networks, I show that diatoms act as repulsive segregators in the ocean, in particular towards potentially harmful organisms such as predators as well as parasites, and that species co-occurrence is driven by environmental factors in a minority of cases. By leveraging the singularity of the *Tara* Oceans data, I provide a comprehensive characterization of a prevalent biotic interaction between a diatom and heterotrophic ciliates at large spatial scale, illustrating the success of data-driven research. Overall, my thesis contributes to our understanding of diatom biotic interactions, from the global patterns to the single cell.

Mots clés (français) :
Plancton, diatomées, biodiversité marine, interactions microbiennes, réseaux de corrélation microbiens, structure des communautés, cellule unique, biologie écosystémique

Keywords :
Plankton, diatoms, marine biodiversity, microbial interactions, microbial correlation networks, community structure, single cell, ecosystem biology