Ecology of Marine Diatoms through Omics: From Community Structure to Single Species Investigation

Thesis

Ecology of marine diatoms through omics:

from community structure to single species investigation

Lucia Campese

Thesis submitted for the degree of Doctor of Philosophy in Life, Health and Chemical Sciences

March 2022

To my mother

for teaching me the importance of

studying, understanding and knowing

Abstract

Diatoms are a major component of marine phytoplankton and play a key role in global elemental cycles and marine food webs. Their ecological success relies to their underlying genetic and functional diversity, both at community and single species levels; despite their importance, diatoms' biology and ecology is far from being completely understood, given the scarcity of global-scale observations and the large number of uncultured species. The general aim of my thesis was to explore diatom global-scale diversity from community to species level by applying and developing different methodological frameworks. This study was mainly based on metabarcoding, metagenomic and metatranscriptomic data sets provided by *Tara* Oceans and *Tara* Oceans Polar Circle expeditions. I first described diatom communities from a statistical perspective to provide a general overview of the macroecological structure of different types of High Throughput Sequencing data. I further investigated diatom diversity across the currents that bring waters from the North Atlantic to the Arctic Ocean; I here characterized taxonomic composition of communities, along with the description of the main environmental features that shape diatom assemblages. Moreover, I described how the expression of functional genes involved in iron uptake, transport and varied across the sampling sites. I then moved to a lower taxonomic rank and focused on a single genus, i.e., *Pseudo-nitzschia*, one of the most abundant and ubiquitous diatom genera that includes toxigenic species responsible of harmful blooms. I described *Pseudo-nitzschia* biogeography at global scale at high taxonomical resolution and explored its ecology through links to abiotic and biotic factors, with a particular focus on the interactions with bacteria. Finally, I performed a functional analysis of a set of genes whose expression in diatoms is strongly up-regulated during sexual reproduction; by integrating a phylogeny-based approach, I tested the role of these genes as potential molecular markers to detect sexual reproduction events in diatom natural populations.

Acknowledgements

I thank my Director of Studies, Daniele Iudicone, for pushing me beyond my limits and for widening my biologist mind-set by trying to make me see the ocean from an oceanographer perspective.

I thank my external supervisor, Olivier Jaillon, for having warmly welcome me in Paris during my first stay, and for introducing me to colleagues at Genoscope.

I thank Maurizio Ribera D'Alcalà and Marina Montresor for their constructive feedbacks on my thesis. I thank Roberta Piredda and Valeria Ruggiero for their help on metabarcoding analysis and Mariella Ferrante for involving me in her projects.

I thank the “extended" Procaccini lab, including the ex-members and the new ones, and in particular Miriam, Manu, Domenico, Lazaro, Chiara e Daniele: it always feels like home to be with you. A special mention goes to Gabriele, an example of strength, intelligence, sensitivity, empathy, generosity, honesty. I am very grateful to have met you and I learnt a lot from you.

I profoundly thank Bruno HayMele for his invaluable help on countless issues, from dealing with R scripts to less ordinary subjects.

I also want to thank Mariano, Luca and Vincenzo from SZN and Kevin and Jana from Genoscope for their friendship during these years. A special mention goes to colleagues met in these years that became essential friends for me: Romuald, Lorenzo and Camilla.

I thank my lifelong friends, Magda, Federica, Nadia, Stefano, Jude, Martina, Clara, and in particular Uvi and Claudia, for their unfailing support and love.

I thank my family: my mother, my father, Lollo, Bimbo e Bullo, for having lightened so many hard moments during these years and for taking care of me.

Finally, I thank Alessandro, for having constantly accompanied me over these years, often being the only one able to pull me away from the “event horizon”.

Summary

List of Figures

List of Tables

Chapter I

Background

1.1 Diatoms

The oceans occupy nearly 71% of the planet's surface and represent the largest continuous environment on Earth. Oceanic ecosystems emerged about 3.5 billion years ago, developing complex trophic interactions encompassing millions of species (*Falkowski and Knoll, 2007*). These ecosystems are nowadays being transformed by anthropogenic climate change, whose imprint is already evident on multiple ocean properties (*Dore et al., 2009; Allen et al., 2018*) that shape the physiology and ecology of marine life (*Boyd et al., 2018*).

Plankton, and mainly phytoplankton, responsible for about half of total primary production on Earth, supports life in this vast environment and represents a key component in the functioning of the global biogeochemical cycles (*Field et al., 1998; Falkowski et al., 2004*). Plankton communities are complex systems whose structuring is strictly linked to both abiotic and biotic components of the oceanic environment; while interacting with other organisms, plankton drive and respond to rapid fluctuations in temperature and nutrients as well as longer term variations (*Hays et al., 2005; Doney et al., 2012*); understanding this response to hydrographical and meteorological forcing is crucial in the present context of anthropogenic climate change (*Estrada et al., 2016*).

Of key importance within plankton communities are diatoms (Bacillariophyta), unicellular photosynthetic microalgae belonging to the clade Stramenopiles (also known as Heterokonta). Diatoms represent a major class of phytoplankton in both marine and freshwater environments (*Field et al., 1998; Kooistra et al., 2007*) and are responsible for about 25% of the total carbon fixed on Earth (*Nelson et al., 1995; Field*

et al., 1998), more than all rainforests combined, producing one fifth of the oxygen we breathe. Therefore, they have a major ecological significance and impact on the global elemental cycles (*Ragueneau et al., 2000; Tréguer, 2002; 2018; Jin et al., 2006*) and in maintaining the ecosystem stability.

Diatoms are the most diverse unicellular microalgae, holding a number of species ranging from 12,000 to 30,000 according to recent estimates (*Guiry, 2012; Mann and Vanormelingen, 2013; Malviya et al., 2016*). They are ubiquitous, their distribution extends from tropical and subtropical regions to polar ecosystems, and are dominant in regions of high productivity (upwelling zones) and high latitudes (the Arctic Ocean and the Southern Ocean). Such richness of species is reflected in a huge variety of shapes, sizes, lifestyles and survival strategies that are expression of their ability to effectively adapt to the highly dynamic marine environment, making diatoms one of the most ecologically successful groups among photosynthetic eukaryotes (*Vanormelingen et al., 2007*).

According to phylogenetic studies and evidences provided by fossil records, diatoms arose around the Triassic-Jurassic boundary, about 190 million years ago (*Nakov et al., 2018; Falciatore et al., 2020*), and have since then undergone extensive adaptive radiation (*Godhe and Rynearson, 2017*). Diatoms derived from a secondary endosymbiosis in which a heterotrophic eukaryote host engulfed a red alga that had previously been created by a primary endosymbiosis between a cyanobacterium and a phagotrophic eukaryote (*Moustafa et al., 2009*).

The most distinctive morphological feature of diatoms is their elaborate cell wall known as the frustule, made of amorphous silica and comprising two valves (thecae) that fit together like a Petri dish (*Round et al., 1990*). Subcellular organelles include the nucleus and a plastid surrounded by four membranes and connected to a rough endoplasmic reticulum. The frustule defines and maintains cell shape and helps protect cells against grazers and pathogens, yet imposing restrictions on mechanisms associated with the life cycle, i.e., cell division and expansion. Moreover, frustule dissolution rate influences the sinking behaviour of diatoms

from sunlit upper layers to the ocean interior, a process that, in turn, removes carbon from the global carbon cycle over geological time scales (*Falkowski, 1997*), ultimately increasing the amount of oil and gas fossil fuel reserves. Moreover, the mineralization of silicon by diatoms is crucial step of the silica biogeochemical cycle (*Tréguer and De La Rocha, 2013*). Frustules show a wide diversity of forms, that are the basis of the traditional diatom taxonomic classification into two main orders: the radially symmetrical centrics (Centrales) and the bilateral symmetric pennates (Pennales) (*Round et al., 1990, Kooistra and Medlin, 1996*). Centric diatoms are further divided into the evolutionary most ancestral radial centrics (Coscinodiscophyceae) and the polar centrics (Mediophyceae), while pennate diatoms can be subdivided into araphid (Fragilariophyceae) and raphid species (Bacillariophyceae). The latter (e.g., *Phaeodactylum tricornutum*, *Fragilariopsis cylindrus*, *Pseudo-nitzschia multistriata*) represent the most recent group (*Kooistra et al., 2007*) and are characterized by the presence of the raphe, a longitudinal slit in the cell wall involved in cell motility (*Poulsen et al., 1999*).

Diatom diversity is not limited to their external morphology, but is also reflected in their ability to live both as free-floating single cells and in differentially shaped colonies (e.g., flat, spiral, ribbons, fans or star shaped) and in their variable size: cell volume spans almost 9 orders of magnitude with the larger cells exceeding 10^9 μm^3 (*Litchman et al., 2009*). Their metabolic strategies range from photoautotrophy to heterotrophy, also including parasitism, like in the case of parasitic diatoms found inside Antarctic sponges (*Bavestrello et al., 2000*). Moreover, evidence of facultative mixotrophy has been found in some species of *Nitzschia* (*Kitano et al., 1997*), in *Skeletonema costatum* (*Guihéneuf et al., 2008*), *Phaeodactylum tricornutum* (*Villanova et al., 2017*) and *Thalassiosira pseudonana* (*Baldisserotto et al., 2021*); in particular, mixotrophy has been proposed as a metabolic strategy for diatoms to face the harsh winters at high latitudes, where low temperature, short photoperiod and low irradiation create unfavourable conditions for photosynthesis (*Villanova and Spetea, 2021*).

As stated before, diatom's successful adaptation to the ocean's fluctuating environment is evident both from their ubiquity, which reflects their ability to survive in stratified oligotrophic waters in suboptimal environmental conditions, and from their dominance during periodic and highly diverse bloom episodes (*Guillard and Kilham,1977*). When non abundant, diatoms are believed to survive in deeper layers thanks to buoyancy regulation, or as quiescent stages. In particular, many species can produce dormant stages that could act as a seed bank for future years when resuspended in the sunlit layer of the water column; this reservoir of diversity is an essential asset that ensures an overall plasticity of response of the whole community to environmental variability (*Malviya et al., 2016*). Diatoms are then ready to take advantage of favourable ecological conditions as and when they arise, as observable from bloom episodes. These events are strongly linked to seasonality, being essentially driven by light and nutrient availability, but also to biotic interactions including predation and competition (*Falciatore et al., 2020, and references therein*). Diatom blooms, and particularly spring blooms in eutrophic coastal regions, are primary productivity phenomena in marine systems that influence both higher trophic levels and biogeochemical export fluxes. During these events, diatom species alternate according to their size and ecological requirements (*Margalef, 1978; Guillard and Kilham, 1977*): the bloom starts after upwelling or strong mixing, and it is originally dominated by fast-growing species belonging to the *Thalassiosira*, *Chaetoceros*, or *Skeletonema* genera; at the end of the bloom, as nutrients are progressively being consumed, large diatoms that are more adapted to oligotrophic environments appear, such as *Rhizosolenia* and *Hemiaulus* species, often associated with nitrogen-fixing cyanobacteria. However, exceptions to this typical three-steps diatom succession have been documented, and a recent study (*Leblanc et al., 2018*) stated that nanoplanktonic diatoms (<5 µm), like *Minidiscus* species, can be major contributors to spring blooms both in coastal and open ocean regions, and advocated for a major consideration of these overlooked species.

The high proliferation rate during blooms is driven by rapid mitotic divisions, processes during which the frustule undergoes characteristic changes (*Round et al., 1990*). For many diatoms, each cell division forces each daughter cell to synthesize a new hypotheca (the smaller valve) inside the theca inherited by the mother cell (*Macdonald, 1869; Pfitzer, 1869*), leading to a progressive reduction in cell size as long as divisions occur. Not all diatoms experience this shrinking of cells dimensions, as attested by the cases of *T. pseudonana* and *P. tricornutum* (*Montresor et al. 2016*). For the ones that have to cope with this decrease in mean size of daughter cells, two ways to restore the original size are vegetative enlargement and sexual reproduction. The latter occurs with different strategies between centric and pennate diatoms, the first showing homothallic reproduction, with eggs and sperm produced in a clonal strain, the second characterized by heterothallism, in which sex occurs only when strains of opposite mating type are mixed together (*Montresor et al., 2016*).

Sexual reproduction has, obviously, not only the role of restoring diatom's cell size, but it is an essential mechanism to establish the genetic structure of a population (*Barton and Charlesworth, 1998; Amato et al., 2005; Kim et al., 2020*) by maintaining the intraspecific genetic diversity via homologous recombination, and ultimately increasing the adaptation potential of local populations to be fit and competitive in their respective habitat (*Godhe and Rynearson, 2017*). However, genetic diversity can be acquired also via mitotic mutations during the vegetative growth phase (*Tesson et al., 2013*) and asexuality can be more advantageous than sexual reproduction in the open ocean, where cell density is scarce and the chance of encountering opposite mating types is low (*Falciatore et al., 2020*). Moreover, other molecular mechanisms such as changes in ploidy have been proposed to play a role in diatom evolution and adaptation (*Parks et al., 2018*); one recent example of these mechanisms is mitotic recombination. A recent study (*Bulánková et al., 2020*) has in fact demonstrated that a high rate of interhomolog mitotic recombination occurs

in diatoms, probably contributing to enhance their plastic response to environmental changes.

All the above-mentioned examples show how the striking macroscopic diversity of diatoms in terms of morphologies, physiologies and life strategies finds its reflection, or more correctly its aetiology, at the molecular level, making diatoms a very interesting taxon to study, even just from a fundamental research perspective.

However, as mentioned before, interest in studying diatoms is also motivated by their involvement in a range of marine ecosystem services that ultimately affects human life and society. Cited examples are their role in the main biogeochemical cycles as well as in blooms, a phenomenon that in turn supports many of the world's most productive fisheries. Diatom blooms can also be dangerous; this phenomenon occurs when those proliferations, called HABs (Harmful Algal Blooms) are associated with algae-produced toxins that have deleterious effects on shellfish, fish, birds, marine mammals and ultimately humans, like the case of several *Pseudo-nitzschia* species (*Bates et al., 2018*), or when blooms of non-toxic species reach a certain density in the water that can damage fish gills, as in the documented case of the mortality of salmonids in fish farms due to *Chaetoceros concavicornis* and *C. convolutus* (*Albright et al., 1993*). Predicting and preventing HABs is thus essential for the whole ecosystem and it is nowadays a priority, as climate change is believed to increase their severity and prevalence (*Hennon and Dyhrman, 2020*).

Finally, applied research is increasingly focusing on diatoms for their enormous biotechnological potential as molecular factories to produce therapeutics, biofuels and biopolymers (*Kroger and Poulsen, 2008; Tanaka et al., 2015, Tirichine et al., 2017*). In particular, diatoms are being extensively used for the production of antimicrobial agents and recombinant proteins, as well as nanocarriers used for drug delivery (*Khavari et al., 2021*).

1.2 Diatom Diversity and Distribution

1.2.1 What is Biodiversity

The term "biodiversity" refers to the complexity of life on Earth, and represents the sum of all biotic variations from genes to ecosystems. The underlying mechanisms that lead and maintain biodiversity have long interested ecologists, and their comprehension is crucial to understand evolutionary dynamics, assess ecosystem stability and predict its susceptibility to anthropogenic climate change.

The modern view in evolutionary biology is that random alterations of DNA sequences and chromosomal structure are the primary source of the genotypic variation that, in turn, determines the phenotype (*Laland et al., 2014*). Once this inheritable variation has been created, the environment influences the survival and the reproductive success of the individuals holding different phenotypes, ultimately leading to the Darwinian "survival of the fittest" (*Spencer, 1872*), with the less competitive species undergoing extinction. The notion of competition exclusion, thus already implicit in Darwin's theory of evolution, has been later mathematically illustrated (*Volterra, 1928*) and successively formalized in the Competitive Exclusion Principle (*Gause, 1934; Hardin, 1960*), that states that two species occupying identical niches, i.e., using the same resources and living in the same place, cannot coexist at steady state. There will be one holding even a slight ecological advantage that will eventually out-compete the other and drive it to extinction. According to this view, plankton, and phytoplankton in particular, competing for few resources (i.e., light and nutrients) and living in a relatively homogeneous environment, should be dominated by a small number of species and ultimately reach an equilibrium in which only the fittest species exists. Nevertheless, phytoplankton show a high degree of diversity even in moments of the year in which marine waters are scarce in nutrients and the inter-specific competition is supposed to be at its maximum: how is it thus possible that so many species co-occur in water columns that hold

so few niches? This dilemma, known as "plankton paradox" (*Hutchinson, 1961*), has been questioned by ecologists over the last 60 years and several theories have been conceptualized to explain the reasons of species coexistence in pelagic waters.

One of the first ideas proposed is that the time that elapses between two strong changes in the marine environment is actually always longer than the time needed to for complete competitive exclusion to come to completion, this leading to a "permanent failure to achieve the equilibrium" stated by the Competitive Exclusion Principle (*Hutchinson, 1961*). This role of the environmental heterogeneity has been hypothesized to work both in time (*Robinson and Sandgren, 1983*; *Sommer, 1989*) and in space at micro-scale (*Richerson et al., 1970*) and attributed to weather fluctuations as well as to physical processes like mesoscale oceanic turbulence (*Bracco et al., 2000*). According to niche theory, in fact, if the required resource fluctuates enough in time or space, multiple species can coexist (*Richerson et al., 1970*; *Vandermeer, 1972*; *Tilman, 1982*; *Sommer, 1984*).

Besides abiotic and physical forcing, the role of biotic interactions (e.g., grazing, pathogenicity and parasitism) has also been proposed to explain the plankton paradox. Indeed, top-down mechanisms have been suggested to have the potential for being essential drivers of phytoplankton diversity (*Prowe et al., 2012*; *Lima-Mendez et al., 2015*), and the variety of microalgae shapes and sizes could reflect defence mechanisms against grazing (*Smetacek, 2001*), one example being the diatom frustule (*Hamm et al., 2003*). Moreover, viruses and parasites may infect phytoplankton based on its density, thus mainly affecting the most abundant taxa, limiting the fittest species and allowing other taxa to survive (*Thingstad, 2000*), ultimately increasing phytoplankton diversity ("kill the winner hypothesis"). Moreover, the role of species-specific metabolic strategies has been acknowledged to promote resource partitioning and allow different potential competitor species to coexist even in a homogeneous environment. A recent study on co-occurring diatoms, for example, examined the expression of nitrogen and phosphorous

metabolic genes and proved that coexisting species can modulate their physiology to partition their niche space (*Alexander et al., 2015*). Finally, theoretical studies suggested that the process of resource competition itself leads to chaotic oscillations in communities' structure in terms of fluctuating dynamics of relative abundances, promoting the disequilibrium conditions required for species coexistence (*Huisman and Weissing, 1999, 2000*).

The above-described theories are all collocated in the context of niche theory, that emphasizes deterministic processes based on niche differences between competing species. However, early in *1998, D.A. Siegel* suggested that the conditions that promote the competition we observe in phytoplankton culture experiments are actually rarely verified in nature; according to his study, phytoplankton cells are in fact always too distant between each other to allow for competition with their neighbours on an instantaneous basis, and this is especially true in oligotrophic open ocean environments. *Behrenfeld et al.* (*2021a*) recently described this circumstance as "competition-neutral resource landscape", and suggested that other mechanisms must drive phytoplankton diversity and composition in this condition. In this context, the neutral theory of biodiversity (*Hubbell, 1997, 2001*), inspired by *MacArthur and Wilson*'s theory of island biogeography (*1967*), provides another framework for understanding planktonic communities in the absence of resource-based competition. In Hubbell's model all individuals are believed to hold the same probability to reproduce and die (neutrality concept). The observed variability in relative abundances across species is thus solely consequence of stochastic demographic processes, i.e., birth, death and offspring generation, collectively known as "ecological drift" (*Vellend, 2016*).

Although neutral and niche theories are based on opposite assumptions, recent efforts tried to reconcile them to explain plankton biodiversity. For example, the theory of "lumpy coexistence" proposed that species are self-organized in assemblages that compete with each other, but that, within each assemblage, species hold very similar traits and can be considered as nearly neutral

communities (*Scheffer and Nes, 2006*). Moreover, a recent study (*Behrenfeld et al., 2021b*) proposed the interplay of ecological and neutral processes by assessing that predator-prey interactions are the ecological factor that drives the creation of communities of equally fit species whose biodiversity is maintained by physical mixing and stochastic processes.

But why the maintenance of biodiversity is so important?

There are many reasons why we may want to preserve biodiversity; the most immediate is that biodiversity is what determines the correct functioning of ecosystems, what regulates the Earth and ultimately provides the goods and services to humanity. Many examples from terrestrial ecosystems have shown how intensive land use by humans and the consequent habitat fragmentation and loss of biodiversity can have profound effects on essential properties of ecosystems, making them less resistant and resilient to perturbations and more vulnerable to biological invasions (*Tilman et al., 2014* and references therein). This is particularly relevant in the context of ongoing and future environmental alterations due to anthropogenic climate change, whose effects strongly rely on ecosystem stability. It is clear that a higher biodiversity reflects a higher ability to cope with environmental changes, thanks to a higher probability of having some species actually able to respond to the ongoing changes (*McCann, 2000; Jungblut et al., 2018*), or thanks to a greater chance that, if a species is lost, a functionally equivalent one takes its place, maintaining the ecosystem stability over time (*Yachi and Loreau, 1999; Jungblut et al., 2018*). In this context, it is more the functional traits of species than their taxonomic identity that actually determines the persistence of biological and biogeochemical processes within an ecosystem (*Diaz et al., 2006; Jungblut et al., 2018*), and functional information in predictive studies on ecosystems has been progressively integrated (*Loreau et al., 2001*, and references therein). Functional traits are components of the phenotype of an organism that influence its fitness: they can be morphological or physiological characteristics or a life or trophic strategy, or features that regulate the resource acquisition, the reproduction and

the overall survival of an organism (*Violle et al., 2007*; *Litchman and Klausmeier, 2008*; *Litchman et al., 2013*; *Ramond et al., 2019*). Species showing similar functional traits are thought to adopt similar ecological strategies and play similar functional roles within ecosystems (*Diaz et al., 2013*).

Regarding phytoplankton, their functional properties in the context of ecosystem functioning have been largely incorporated in theoretical and model studies on biogeochemical cycles, as different groups differentially influence the global cycles of carbon, nitrogen and silicon (*Le Quéré et al., 2005*; *Litchman et al., 2015*). These models usually categorize the biodiversity of phytoplankton into a limited number of plankton functional types (PFTs) that are indicative of key biogeochemical functions (*Le Quéré et al., 2005*; *Tréguer et al., 2018*), and whose parametrization is based on experimental as well as observational studies (*Litchman et al., 2015*, and references therein). In this context, diatoms are identified by a unique PFT, according to the historical view of them as r-strategist life-forms characterized by rapid growth and highly favoured in high turbulence and high nutrients concentrations (*Margalef, 1978*). However, recent studies have provided valid attempts to reduce the oversimplification of gathering the high variability of diatoms into a single functional type through different approaches (all reviewed in *Kemp and Villareal, 2018*), such as the use of two diatom PFTs (*Stukel et al., 2014*), the account for functional diversity within the existing diatom PFT (*Terseleer et al., 2014*) and the development of more flexible PFT models (*Smith et al., 2016*). To conclude, it is now clear that diatoms cannot be considered as an individual PFT and the implications of this conclusion are particularly important in biogeochemical modelling studies that attempt to predict how ocean ecosystem will respond to climate change (*Kemp and Villareal, 2018*).

1.2.2 Measures of Diversity and Community Structure

As described in the previous section, biodiversity is a complex concept that can be expressed at every level of biological organization. Communities, defined as groups of species co-existing in space and time (*Fauth et al., 1996*), are often the hierarchical levels used to address questions regarding spatial and temporal variation of biodiversity (*Magurran, 2004*). With the advent of Next-Generation Sequencing techniques and the introduction of Operational Taxonomic Units (OTUs) as a proxy for species, together with the development of functional diversity concepts, measurements of biodiversity are being progressively applied to biological units other than the traditional "biological species" one.

Given its complexity, there is no unique way to effectively quantify biodiversity in an exhaustive way, and a wide array of measures and metrics has been developed (*Magurran, 2004*). However, when measuring community structure and diversity two main components are to be taken into account: the number of different taxonomic or functional units considered, and their relative abundances, respectively richness and evenness. Richness simply indicates the total number of species in the ecological sample. Being at the same time very informative and intuitive, it is by far among the most used descriptors in community ecology (*Magurran, 2004*), with applications that range from community ecology to ecosystem conservation. Evenness reflects how the relative abundances of species are distributed within a community. High evenness means that all species in a community have the same abundance, while the opposite can reflect the common condition in which there are a few very abundant species and many species are rare. In some way inverse of evenness is the dominance index, that refers to how much one or a few species numerically dominate the community (*McNaughton and Wolf, 1970*).

Two widely used measures of diversity in ecology that account for both richness and evenness are the Shannon (*Shannon, 1948*) and the Gini-Simpson indices (*Simpson, 1949*). The first weights species according to their proportional

abundance, while the latter represents the probability that, randomly picking two individuals in a community, they belong to different species. Richness, Shannon and Simpson indices differ in the emphasis given to rare species: the first, ignoring abundance information, emphasizes the rare component of a community, the second considers the relative abundance of the single species, while Simpson emphasizes the dominant ones (*Chao and Jost, 2012*; *Pallman et al., 2012*).

Nevertheless, there is no single index able to capture the complexity of communities. An alternative is to examine the structure of a community by analysing the distribution of species abundances (SAD: Species Abundance Distribution) of the (sampled) community, which provides more detailed information and thus a more integrated description (*Magurran, 2004*; *McGill et al., 2007*). A community SAD displays the abundance for each species within a community and is normally plotted in the form of a histogram of absolute number of species on the y-axis and their abundance on an x-axis. The classic hyperbolic shape of this visualization, often referred to as the "hollow curve" clearly indicates how the majority of communities are uneven, with few very abundant species and many rare species (*Fisher et al., 1943*; *Preston, 1948, 1962*; *Whittaker, 1965*; *Brown, 1995*; *Gaston and Blackburn, 2000*). SADs are also often displayed through a rank-abundance distribution (RAD) diagram, where log-abundance is plotted on the y-axis and rank on the x-axis. By accounting for abundances, SADs thus allow to assess the difference in the degree of evenness, dominance or rarity between communities. Moreover, for construction, SADs only show occurring species, while not considering the "zero abundance" entities; this feature allows the comparison between communities that have little or no taxonomic overlap, e.g., a diatom and a forest community. This characteristic, together with their intuitive visual nature, made SADs very popular in ecological research (*McGill et al., 2007*). SADs have been largely studied for plankton, with the common outcome being the presence of high numbers of rare species in these communities (*Sogin et al., 2006*; *Caron and Countway, 2009*; *Ser-Giacomi et al., 2018*).

Richness, evenness, dominance, rarity and SADs are all descriptors of communities that measure the diversity at a single site, and are thus included in what is known as the α-diversity component of biodiversity. They do not provide any information of species composition; hence, in order to draw conclusions on the compositional variation between two or more sites, other metrics are required, all ascribed to the category of β-diversity (*Whittaker, 1960*). The sum of the α and β diversity gives γ-diversity, i.e., the component that describes the overall variation of a landscape or region. As with α-diversity, a large number of β-diversity indices has been proposed (*Tuomisto, 2010a, b*; *Anderson et al., 2011*). In general, β-diversity is considered as the opposite of similarity, conceptually reflecting the compositional heterogeneity among sites, and can be related to different aspects of community composition similarity (*Koleff et al., 2003*; *Baselga, 2010*; *Tuomisto, 2010a*; *Anderson et al., 2011*, *Antão, 2019*). Compositional similarity can be measured by various indices that are calculated using presence/absence or abundance data: they usually range from 0 to 1, with 0 indicating that the two communities have no species in common, and 1 reflecting complete overlap in species identities.

Within this thesis, I will provide an overview on diatom communities inhabiting the global epipelagic ocean by studying their SADs and measuring alpha diversity indices like richness, evenness, dominance and rarity. I will apply this approach to both taxonomic and functional units, with the aim of revealing the emergent properties of diatom community structure and discuss the implications of the observed patterns. Moreover, I will use an integrated approach to characterize the compositional difference of communities along an oceanographic transect by applying several β-diversity indices in the context of major temperature, salinity and nutrient environmental gradients.

1.2.3 Diatom Distribution

As described in section 1.2.1, biodiversity patterns are the result of coexistence of species in distinct regions of the globe. The investigation of biodiversity and its determinants is thus inherently linked to biogeography, i.e., the study of the geographic distribution of species, and the variation in the extent of their spatial ranges.

For terrestrial species, usually characterized by low dispersal capacity, a strong relationship between geographic and genetic distance has been shown (*Carr et al., 2003*). In contrast, marine planktonic microorganisms, immersed in a global-scale fluid environment, are thought to have a huge dispersal potential allowing high gene flow, large and homogenous populations, and little divergence over even large spatial scales (*Whittaker et al., 2017*). This concept can be summarized by the tenet of "everything is everywhere but the environment selects" ("global ubiquity hypothesis", *Baas-Becking, 1934*), which became a paradigm in microbial ecology (*Fenchel et al., 1997*; *Finlay, 2002*; *Azovsky, 2000, 2002*). However, despite the evolutionary significance of environmental selection, quantifying its role in determining population and community structure in the global ocean remains challenging. The development of the neutral model of biodiversity (*Hubbell, 1997, 2001*) has added complexity to the debate. As described in section 1.2.1., in Hubbell's model all variability is due to demographic stochasticity, which gives new value to the role of dispersal and geography in structuring populations.

The debate on the importance of dispersal-driven and environmental-driven processes in assembling communities is particularly relevant to phytoplankton; in fact, despite their enormous dispersal potential which should homogenize their gene pools, phytoplankton species show evidence for highly structured populations and communities (*Rynearson and Ambrust, 2000*; *Iglesias-Rodriguez et al., 2006*; *Casteleyn et al., 2010*; *Erdner et al., 2011*; *Godhe et al., 2016*). In most marine ecological systems, quantifying the relative contribution of niche and neutral

processes in determining biogeographical patterns remains a challenge (*Gilbert and Lechowicz, 2004*; *Watson et al., 2011*). However, during the last two decades, many studies have emphasized the importance of both processes in shaping the distribution of phytoplankton communities, demonstrating that niche and the neutral paradigms are not contradictory, but rather two complementary extreme views of what actually occurs in nature (*Barton et al., 2010*; *Chust et al., 2013*; *Estrada et al., 2016*; *Monchamp et al., 2018*).

Diatoms' ubiquity and abundance make them useful model organisms for exploring the important links between dispersal, environmental selection and organism evolution. In a study performed with genetic markers at local scale and at species level, *Godhe et al. (2013)* showed that the patterns of genetic structure on the centric diatom *Skeletonema marinoi* could not be explained by genetic models based on Isolation-By-Distance (IBD; *Wright, 1949*), but were best described by local oceanographic connectivity. Hence, if at global scales dispersal may reflect the Euclidean distance and lead to classic IBD patterns, this may fail at regional level, where local mixing and advection can complicate the oceanographic circulation and lead to connectivity patterns that cannot be solely explained by geography (*Treml et al., 2008*). Mixing and dispersal can influence plankton distribution indirectly, via changes in nutrient concentrations, or directly, when together with advection of water masses introduce species to inhospitable regions contributing to the creation of source-sink dynamics (*Beaugrand et al., 2007*; *Villar et al., 2015*). Taking into account ocean connectivity is thus essential when studying plankton ecology and biogeography, both at regional and local scale and at community and molecular level. Similar results on *S. marinoi* were obtained by *Sjoqvist et al. (2015)*, who explained the genetic differentiation patterns in the North Sea–Baltic Sea transition zone as the result of oceanographic connectivity patterns creating an asymmetric migration between the two basins which supports local salinity adaptation.

The biogeographic structure of a population at a global scale has been first investigated in the pennate diatom *Pseudo-nitzschia pungens*; here geographic distance appeared to be a strong barrier to gene flow, suggesting allopatric isolation despite high dispersal potential (*Casteleyn et al., 2010*). The potential for biogeographic patterns deriving from neutral evolution has further been demonstrated in modelling efforts on marine bacteria (*Hellweger et al., 2014*). In contrast, some studies of centric diatoms suggested that population structure may not necessarily be dominated by allopatric processes (*Godhe and Härnström, 2010*; *Rynearson et al., 2006*). *Whittaker and Rynearson (2017)* used microsatellite markers to identify genetically distinct populations within the globally distributed, bloom-forming diatom species *Thalassiosira rotula*. They found no correlation between geographic distance and population structure on global scales, but still observed distinct populations in this species, concluding that the divergence observed was a product of ongoing selective mechanisms that allow populations to persist over time frames greater than decadal-scale global surface seawater connectivity and perhaps even longer.

The ongoing growth of increasingly finer molecular investigations has provided further insight on genetic structure within and among globally distributed species. In particular, recent studies led to the discovery of several cryptic – morphologically indistinguishable but genetically different – diatom species, as in the case of the *Chaetoceros socialis* species complex (*Chamnansinp et al., 2013*; *Gaonkar et al., 2017, 2018*; *De Luca et al., 2019*) and the genera *Skeletonema* (*Sarno et al., 2005, 2007*; *Zingone et al., 2005*) and *Leptocylindrus* (*Nanjappa et al., 2013*). Studies on genetic structuring among large numbers of strains belonging to these genera have shown that geographical distribution patterns differ markedly among groups of closely related species (*Kooistra et al., 2008*; *Casteleyn et al., 2010*), suggesting that cryptic diversity is common in diatoms, where lineage sorting, as deduced from genetic data, proceeds faster than morphological differentiation (*Alverson, 2008*).

The finding of remarkable cryptic diversity along wide spatial scales again questions the global ubiquity hypothesis and suggests that previously recognized cosmopolitan species are actually composed of distinct entities (*Rengefors et al., 2017*; *Amato et al., 2018*). Increasing information derived from metabarcoding and metagenomics approaches, together with data on seascape configuration, could provide additional knowledge for understanding the biogeography of diatoms and the level of genetic diversity of distinct populations and species (*Amato et al., 2018*). Understanding diatom biogeography and the degree of genetic diversity of its populations is particularly important since these organisms are the dominant phytoplankton group at high latitudes, regions particularly affected by climate change (*Meredith et al., 2019*), and overall constitute a relevant fraction of the oceanic primary producers at global-scale.

It is known from Earth system models that the future ocean will be warmer, more acidified, less oxygenated, more stratified at the surface and less mixed at depth (*Steinacher et al., 2010*; *Bopp et al., 2013*; *Barton et al., 2016*). Plankton species are expected to respond to these changes by altering the internal structure of their communities, by shifting their biogeographical patterns or by adaptively evolving to the new environment (*Hutchins and Fu, 2017*; *Abirami et al., 2021*). These three main mechanisms are not mutually exclusive and their interplay depends on multiple conditions, from the degree of the environmental changes to the establishment of biotic interactions. A possible aspect of community reshuffle consists in the variation of species richness, which is expected to generally decrease in the subtropical gyre but increase at high latitudes, as estimated by *Thomas et al. (2012)* and further predicted for diatoms (*Barton et al., 2016*; *Busseni et al., 2020*). The same pattern, although not confirmed for phytoplankton, has been predicted for zooplankton by a recent modeling effort (*Benedetti et al., 2021*). Richness changes are strongly linked to the migration of organisms, that in turn alters their biogeographical patterns. Climate-change-driven shifts in the spatial distribution of plankton are already a fact (*Hallegraeff et al., 2010*; *Poloczanska et al., 2013*); temperate

dinoflagellates and coccolithophores have extended their biogeographic boundaries to higher latitudes (*Hallegraeff et al., 2010*; *Winter et al., 2014*), and diatoms are also predicted to move northward (*Bopp et al., 2005*), at a rate higher than previously reported (*Barton et al., 2016*). However, both communities reshuffle and their shifts cannot be looked at without accounting for the evolutionary component: diatoms are plastic organisms that can rapidly adapt and acclimate to new environmental conditions through different molecular mechanisms whose complexity is still challenging the scientific community. The plasticity of diatoms' environmental niche has been demonstrated relatively to temperature both by observational (*Chivers et al., 2017*) and experimental (*Jin and Agusti, 2018*) studies, the latter also highlighting the forced trade-offs caused by the thermal adaptation. However, temperature is not the only variable impacted by climate change, and recent studies pointed to the role of variation in surface currents in shaping frontal structures together with light, salinity and nutrients patterns that, in turn, are major drivers of plankton biogeography (*Barton et al., 2016*; *Oziel et al., 2020*).

In this context, a main objective of the present thesis has been to study diatom communities along the North Atlantic – Arctic Ocean current, that represents the main oceanic pathway along which the predicted invasion of the Arctic Ocean by species from lower latitudes would occur. By accounting for the high diversity of diatoms, I investigated the taxonomic composition of organisms that are transported along the current and considered how the environmental landscape shapes the observed diversity in this highly vulnerable region.

1.2.4 Diatom Role in Biogeochemical Cycles

The ecological success of diatoms, reflected in their ubiquity and dominance in ocean ecosystems, make them crucial contributors to global biogeochemical cycles. Diatoms play indeed a fundamental role in the carbon cycle (*Smetacek, 1999*) as well as in nutrient cycling of nitrogen and silicon (*Nelson et al., 1995*; *Armbrust,*

2009; Bowler et al., 2010; Tréguer and De La Rocha, 2013); their metabolic activity is also strongly linked to the global cycling of trace elements like iron, cobalt or zinc (Hutchins and Boyd, 2016; Taylor and Sullivan, 2008; Vance et al., 2017). As a general mechanism, diatoms' participation in biogeochemical cycles starts from the acquisition of the element from the environment and its following introduction in the food web through higher trophic levels or its sedimentation at the bottom of the sea floor through sinking processes.

For the carbon cycle, diatoms, as well as the other members of phytoplankton, integrate CO_2 to produce organic matter via photosynthesis, a process that occurs in the euphotic zone and fixes the dissolved inorganic carbon into its organic forms. Successively, phytoplankton is grazed by micro and macro zooplankton or indirectly consumed by heterotrophic bacteria, which will in turn be eaten by larger organisms. The majority of carbon is thus remineralised to inorganic CO_2 through respiration, but a small part escapes consumption. During this process detrital matters, faecal pellets and dead cells are produced, a marine stock that can undergo resuspension in the water column or sink at the bottom of sea floor, allowing for sequestration from the atmosphere over geological timescale (Menon et al., 2007; Ciais et al., 2014). The overall mechanisms of CO_2 acquisition, transformation into organic carbon, sinking and decomposition at depth, is known as the biological carbon pump (Volk and Hoffert, 1985), and represents the most important biological process in the Earth system for the removal of CO_2 from the carbon cycle, a process in which diatoms, by fixing the amount of carbon of all the terrestrial rainforests combined (Field et al., 1998), constitute essential players. Strongly linked to the carbon cycle is the silica cycle (Pondaven et al., 2000), in which diatom contribution is considered predominant, at least in coastal zones where diatoms precipitate circa 240 Tmol of silica per year (Tréguer and De La Rocha, 2013). As in the case of carbon, silica is produced in the euphotic zone and can undergo recycling in the surface layer or sedimentation in the deep ocean. However, not all diatom species play the same role in the global carbon and silica cycle and recent

studies conceptually summarized this differential behaviour by subdividing diatoms into two functional groups (*Assmy et al., 2013*; *Quéguiner, 2013*; *Tréguer et al., 2018*): The C-sinkers and the Si-sinkers. The first consists of small and fast-growing species that tend to form chains, like the genera *Chaetoceros* and *Pseudo-nitzschia*, that dominate iron-replete regions of the ocean; the second is represented by large and slow-growing diatoms holding a heavier siliceous frustule that helps them to avoid grazing, like the species *Fragilariopsis kerguelensis* and *Thalassiothrix antarctica*, mainly found in regions of iron limitation.

Another fundamental component of ecosystem functioning is nitrogen, that after hydrogen, oxygen and carbon represents the fourth most abundant element in organic matter. Nitrogen is often considered a limiting element in the ocean, together with phosphorous, and its cycle is pivotal in the overall ocean biogeochemistry (*Zehr and Kudela, 2011*). Dissolved dinitrogen gas (N_2) is primarily integrated in the food web through the activity of the N-fixers, i.e., bacteria and archaea that can reduce this molecule and release nitrogen in its dissolved inorganic forms like nitrate (NO_3^-) nitrite (NO_2^-) and ammonium (NH_4^+). Nitrate is the most chemically stable form of nitrogen and, together with ammonium, is the main nitrogen source for diatoms and phytoplankton in general. NO_3^- and NH_4^+ uptake by phytoplankton is normally followed by the integration in the trophic network through grazing and microbial decomposition activity, and by the sedimentation to the bottom of the sea.

Along with carbon, silicon, nitrogen and phosphorus, diatom metabolic activity requires a suite of essential micronutrients that are present at trace concentrations (<0.1 μM) and whose abundances in the ocean are strongly linked to ecosystem functioning both directly and indirectly, i.e., with connections to the biogeochemical cycles of main macronutrients (*Morel and Price, 2003*). In particular, these elements play important biological roles both as essential cofactors of enzymes and as structural components of proteins, like the cases of Mn, Fe, Co, Ni, Cu, Zn and Cd. Although occurring in high concentrations within rocks and soil,

trace metals are scarce in the sea mainly because of their limited solubility and their effective export from the water column (*Morel and Price, 2003*); in fact, plankton organisms require trace metals to survive and developed several mechanisms for their uptake, with the overall effect that these elements (with the notable exception of Mn) are depleted at the sea surface. As for macronutrients, micronutrients acquired by diatoms and coccolithophores undergo a downward flux partially balanced by the remineralisation that occurs at depth by heterotrophic bacteria (*Morel and Price, 2003*). Among trace elements, iron is undoubtedly the most studied, given its profound influence of ocean productivity patterns (*Boyd and Ellwood, 2010*). For this element, two main conditions have been described (*Tagliabue et al., 2017*): at high latitudes, where diatoms dominate the phytoplankton community, the main processes that bring iron in the surface waters are upwelling and mixing, with the following recycling by zooplankton and bacterial activity, while at low latitudes an important role is provided by dust and its associated iron supply that enhances the activity of diazotrophs (N-fixers). Iron speciation, its concentration and the number of molecules that bind it have been identified as essential factors to consider when studying the complex role of this micronutrient in the context of global ocean biogeochemistry (*Tagliabue et al., 2017*); moreover, the contribution of iron to carbon and nitrogen cycles is strongly related to its uptake and assimilation by the biological compartment of ecosystems. Recent studies have shown that phytoplankton species (*Behnke and LaRoche, 2020*), and diatoms in particular (*Gao et al., 2021*) display a wide array of strategies for iron acquisition, being able to access different forms of this element in the ocean. Iron transformations are catalysed by microorganisms; combining measures of nutrient concentration with information derived by the biological activity is thus an approach that provides powerful insight into both the understanding of nutrient cycles and the ecological and evolutionary forces that shape biological response of organisms to environmental factors. The investigation of these mechanisms is thus nowadays crucial to understanding how ocean productivity will trend in an ocean expected to be characterized by a decreased primary production and less carbon sinking (*Bopp et al., 2005*; *Stock et al.,*

2014; Barton et al., 2016). In this context, in my thesis I will explore the variation of iron metabolism in diatoms along the oceanic current that brings seawater from the subpolar regions of the North Atlantic into the Arctic Ocean through an approach based on the expression of genes involved in the metabolism of this essential element.

1.3 *Pseudo-nitzschia*

1.3.1 Biogeography and Bloom Dynamics

Pseudo-nitzschia H. Peragallo (1899) is one of the most common diatom genera. Its species are bilaterally symmetrical pennate diatoms quite easily recognizable at microscope; they belong to the raphid group, since their frustules contain a central raphe that, by secreting an extracellular mucilage mechanically coupled to the actin-myosin cytoskeleton across the plasma membrane, allows the cell to move by gliding (Poulsen et al., 1999).

Pseudo-nitzschia species occur in polar, temperate, subtropical and tropical areas worldwide (Lundholm et al., 2002, and references therein), and their cosmopolitan distribution was already assessed through analysis of microscopy data (Hasle, 2002). A recent global-scale diversity description of diatoms using metabarcoding data of all size classes collected within the *Tara* Oceans expedition (Malviya et al., 2016) supported *Pseudo-nitzschia* ubiquity and also found it to rank in the top-ten most abundant genera in the global epipelagic ocean by accounting for 4.4% of the total analysed metabarcoding reads and holding the highest richness in operational taxonomic units (OTUs) among the pennate diatoms. A successive study on a subset of the same dataset used in Malviya et al. (2016), and targeting only the 20-180µm size class, found *Pseudo-nitzschia* to be the second most represented genus

worldwide (after *Chaetoceros*), representing 23% of total OTUs abundance (*Busseni et al., 2020*).

Besides providing new information on species distribution, the use of molecular barcodes has progressively led to the discovery and identification of new species, with the most recent assessment depicting *Pseudo-nitzschia* as one of the most specious diatom genera, holding 58 species (*Dong et al., 2020*; *Guiry and Guiry, 2021*).

Molecular markers also allowed the discovery and identification of species crypticity (see section 1.2.3.) within the genus *Pseudo-nitzschia*. In most cases, the identified cryptic species – morphologically indistinguishable but molecularly different – were phylogenetically close, like the cases of *P. pseudodelicatissima* and *P. cuspidata* (*Lundholm et al., 2012*) and *P. delicatissima* and *P. arenysensis* (*Quijano-Scheggia et al., 2009*). However, in other cases, like the one of *P. arctica* and *P. fryxelliana* (*Percopo et al., 2016*), the morphological identity did not correspond to a phylogenetic proximity of the two species; *P. arctica*, morphologically identical to *P. fryxelliana*, resulted in fact phylogenetically closer to other two species, *P. granii* and *P. subcurvata*, whose morphological features were different from the ones of *P. arctica*. Given the implications that cryptic diversity has for species ecology, these results suggest the importance to integrate different characters (i.e. morphological and molecular) when studying *Pseudo-nitzschia*, a genus for which the timing of morphological divergence, molecular evolution and speciation are often not aligned (*Percopo et al., 2016*).

Moreover, the study of the distribution and structure of some *Pseudo-nitzschia* species shed light on the drivers of diatom biogeography and diversity, which in turn helps understanding the potential for phytoplankton species to evolve and adapt to environmental changes. One example is the case of *P. pungens*, whose global-scale genetic structure was extensively studied using the ITS rDNA marker (*Casteleyn et al., 2010*, *Lim et al., 2014*; *Kim et al., 2018*). *P. pungens* is now believed to exist in the form of three main clades: Clade I (*var. pungens*), Clade II (*var. cingulata*) and Clade III (*var. aveirensis*), that are differentially distributed worldwide and also

show physiological differences in temperature and salinity tolerance (*Kim et al., 2018*). The three groups display a distribution coherent with their geographic origin, thus suggesting the presence of barriers to gene flow between distant locations; these barriers have been supposed to be represented both by patterns created by oceanic currents, as well as land masses (*Casteleyn et al., 2010*; *Lim et al., 2014*). Furthermore, ecological barriers (e.g., differences in water temperature and salinity) have probably contributed to differentiate populations inhabiting regions where gene flow may be permitted by the thermohaline circulation of the global "conveyor belt" (*Lim et al., 2014*; *Bates et al., 2018*). The case of *P. pungens* therefore represents an emblematic demonstration of the importance of the interplay of neutral and niche processes, discussed in section 1.2.3, in structuring unicellular microalgae communities.

Nevertheless, the major interest in *Pseudo-nitzschia* is driven by its capacity to cause Harmful Algal Blooms (HABs, see section 1.1). In particular, it has received much attention since 1987, when a large *Pseudo-nitzschia* bloom occurred in Canada, causing more than a hundred illnesses in humans, and strongly impacting the molluscan shellfish aquaculture industry (*Bates et al., 1998*, and references therein). People affected suffered of various symptoms, and especially showed disorientation and memory loss, after which this clinical syndrome has been termed Amnesic Shellfish Poisoning (ASP; *Bates et al., 1998*). Subsequently, the neurotoxin domoic acid (DA) was identified as the contaminating agent (*Wright et al., 1989*). DA is a water-soluble tricarboxylic amino acid with affinity to glutamate receptors: when it binds them it triggers a massive depolarization of the neurons with subsequent increase in intracellular Ca^{2+} concentration, ultimately leading to neuronal swelling and death. These nerve cells, located in the hippocampus, are functionally linked to memory maintenance, hence the memory loss characteristic of ASP (*Bates et al., 1998*).

Presently, at least 26 *Pseudo-nitzschia* species are thought to be toxigenic; however, not all strains of these species produce detectable concentration of DA and it is

still not clear whether some species that are now labelled as harmless have the potential to release this toxin when properly triggered (*Bates et al., 2018*). Recent progress has been made in depicting the molecular mechanisms of DA biosynthesis (*Brunson et al., 2018*) but it is still not clear which species actually hold the genes involved in DA production and, more importantly, what are the environmental and biological conditions that regulate the activation and inhibition of the expression of those genes.

This is particularly relevant in the context of climate change: a potential link between harmful *Pseudo-nitzschia* and ocean temperature anomalies has been proposed (*McCabe et al., 2016*), consistent with the prediction that a warmer sea is going to host more prevalent and geographically extended HABs. Moreover, although non directly causing HABs, global warming is driving substantial shifts in species ranges; for instance, polar ecosystems become more accessible, and these vulnerable regions must be vigilantly monitored for the presence of toxic diatoms (*Lefebvre et al., 2016*). A more direct example of the role of temperature in promoting HABs is represented by the event that affected the U.S. West Coast in 2015, when a marine heatwave caused an extraordinary toxic *Pseudo-nitzschia* bloom event (*McCabe et al., 2016*) that showed unprecedented spatial extension and duration, along with anomalously high DA quotas recorded (*Bates et al., 2018*, and references therein). The bloom, driven by coastal upwelling, started in mid-April and was initially dominated by *P. australis*, successively replaced by various assemblages of *P. fraudulenta* and *P. delicatissima* complex. Nevertheless, some toxic species were recorded until November in some locations. This bloom affected the whole trophic web, with DA concentration detected in whales, dolphins and sea lions, among others (*McCabe et al., 2016*), and causing huge economic losses. This event occurred during what has been named "The Blob" (*Cavole et al., 2016*; *Di Lorenzo and Mantua, 2016*; *McCabe et al., 2016*), i.e. a warm anomaly that affected Pacific Ocean off the coast of North America from 2013 to 2016. At its peak, parts of this blob reached almost 4 °C above normal water temperature, with an average temperature of 2.5

°C higher than usual. The anomalously warm water is thought to have contributed to the bloom either as a primary factor (*McCabe et al., 2016*) or indirectly, by influencing the upwelling dynamics and in turn the nutrient concentrations (*Ryan et al., 2017*). Indeed, macronutrient conditions are also likely to have played a critical role in triggering the bloom and its toxicity. Nutrient concentrations were ~50% lower than long-term regional means, with an anomalously low silica-to-nitrate ratio (*Ryan et al., 2017*), and nutrient stress has historically been associated to *Pseudo-nitzschia* blooms, like in the case of a toxic *Pseudo-nitzschia* spp. bloom in Alabama waters in summer 2009 (*Liefer et al., 2013*). Moreover, a recent paper (*Clark et al., 2021*) studied dynamics of a highly deleterious bloom that occurred on the opposite side of the U.S. coasts one year after the one triggered by "The Blob". This event interested the Gulf of Maine in 2016 and was again dominated by *P. australis*, a species that was not present there before. While investigating the origin of *P. australis* and the connectivity pathways it could have followed, the researchers also revealed that *P. australis* appearance was associated with a decreased surface seawater salinity.

Abiotic factors and oceanic mixing are not the only features linked to HABs and more precisely to DA production in *Pseudo-nitzschia* species. Indeed, recent studies have emphasized the role of the physical or chemical interaction with zooplankton or bacteria to be implicated in the production of DA (*Lelong et al., 2012*). While the ability of zooplankton grazers to induce DA release has been assessed (e.g., *Tammilehto et al., 2015*), the mechanisms behind the enhanced DA production of *Pseudo-nitzschia* species in presence of bacteria are still to be revealed. The relationships between diatoms and bacteria have been mainly investigated over broad spatial scales; recently, evidence points to the role of microscale interactions occurring within the region immediately surrounding individual phytoplankton cells (*Seymour et al., 2017*). This microenvironment is known as the phycosphere, and represents the planktonic equivalent of the rhizosphere in plants. The exchange of metabolites and chemical cues, including DA, at this interface mediates

phytoplankton–bacteria interactions, which comprise mutualism, commensalism, competition, antagonism and parasitism. A recent study also showed how the bacterial community associated with two *Pseudo-nitzschia* species (*P. hasleana* and *P. mannii*) could have provided them with the ability to degrade polycyclic aromatic hydrocarbons (PAHs) and potentially other environmental contaminants (*Garali et al., 2021*). Host-specific bacterial community has been shown to occur in presence of some *Pseudo-nitzschia* species (*Amin et al., 2015*), but whether these bacteria appear in the phycosphere and/or the surrounding water is unknown. Moreover, host-specificity is normally studied through culture experiment, while the knowledge of what actually happens in nature remains difficult to assess. In conclusion, *Pseudo-nitzschia* blooms are multifaceted phenomena whose aetiology remains unsolved, and given the ubiquity of this important diatom genus and the toxigenic potential of virtually all of its species, it is crucial to better understand what regulates the dynamics of their distribution and growth.

Within this thesis, I will explore *Pseudo-nitzschia* biogeography at global scale with a curated high-resolution metabarcoding dataset, also providing a new assessment of *Pseudo-nitzschia* species abundance and distribution in relation to abiotic and biotic factors, with a special focus on interactions with bacterial communities.

1.3.2 Life Cycle

Pseudo-nitzschia species belong to pennate diatoms and therefore almost all the species are characterized by a heterothallic mating system, where the conjugation of gametes occurs only if they belong to strains of opposite mating types. As discussed in section 1.1, sexual reproduction is a way to restore the normal cell size in diatoms whose rigid silica walls impose cell size reduction during mitosis. In particular, only when reaching a species-specific cell size threshold, a cell acquires the competence for undergoing sexual reproduction and thus restoring the maximum cell size. A study on *P. multistriata* (*Scalco et al., 2014*) showed how not only do cells have to be sufficiently small, but a threshold cell concentration must

also be reached to allow sexual reproduction to start, thus proposing the role of chemical cues to mediate the communication between two opposite mating types, as shown for benthic diatoms (*Sato et al., 2011*; *Gillard et al., 2013*).

The general scheme for the life cycle of *Pseudo-nitzschia* species is based on the detailed information available for *P. multistriata* (*Scalco et al. 2016*): the sexual phase starts with meiotic divisions that lead to the formation of gametes, followed by the conjugation of the haploid gametes of opposite mating types that produces two expandable diploid zygotes that ultimately develop into auxospores (*Davidovich and Bates, 1998*). Within each auxospore, an initial cell of maximum size is formed, permitted by absence of the rigid siliceous cell wall, prior to re-initiating the vegetative phase of the cycle.

Sexual reproduction plays a fundamental role in the life cycle of these pennate diatoms. Although recent progress has been made to uncover the molecular mechanisms behind this process (see *Basu et al., 2017*; *Russo et al, 2018*; *Ferrante et al., 2019*; *Annunziata et al., 2022*), the detection of these events in field population remains challenging. In this thesis I will explore *Pseudo-nitzschia* species sexual reproduction by studying abundance and distribution of several genes that have been recently proposed as putative molecular markers to detect sexual reproduction events in diatoms at sea.

1.4 Environmental Omics: the contribution of *Tara* Oceans

1.4.1 *Tara* Oceans

Ocean complexity has stimulated the curiosity of scientists for a long time and its understanding is a necessity today, since ongoing and future changes due to climate variations are predicted to affect the ecosystem services the ocean provides (*Smith et al., 2021*). It is also clear that a complete understanding of ocean functioning cannot disregard the integration of knowledge of all its physical,

chemical, and biological components, the latter being dominated by planktonic organisms.

Global-scale studies of open ocean organisms date back to the second half of the XIXth century, with the Challenger expedition (1872-1876) that cruised Atlantic, Pacific and Southern Oceans; several other expeditions (e.g., Die Plankton-Expedition, 1889) were successively performed, with a progressive improvement of organism collection and classification. More recently, the advent of new sequencing technologies provided the opportunity to explore plankton worldwide from a molecular ecology perspective, like in the cases of Sorcerer II (*Rusch et al., 2007*), Malaspina 2010 (*Duarte, 2015*) or *Tara* Oceans (*Karsenti et al., 2011*).

In particular, *Tara* Oceans was conceived in 2008 as a multidisciplinary project with the aim of studying planktonic communities on a global scale through the integration of imagery, physicochemical measures and environmental genomics, with standardized protocols. The schooner *Tara* sailed from France in 2009, cruised the Mediterranean and Red Seas and continued through Indian and South Atlantic Oceans, sampled some stations in the Southern Ocean and then went northward through the whole Pacific until the North Atlantic Ocean (Fig. 1.1). Later in 2013, a dedicated campaign extensively sampled the Arctic Ocean (TOPC: *Tara* Oceans Polar Circle expedition). The two expeditions lead to the collection of more than 35,000 plankton samples over 210 stations, for a total of ~13,000 measurements of physicochemical parameters, the generation of 6.8 million images of planktonic organisms and the sequencing of more than 60 terabases of DNA and RNA (*Sunagawa et al., 2020*; Fig. 1.1). The project has been also enriched with further expeditions to assess plastic pollution in the Mediterranean Sea and in European rivers, and to focus on coral reef ecosystems in the Pacific Ocean (*Planes et al., 2019*).

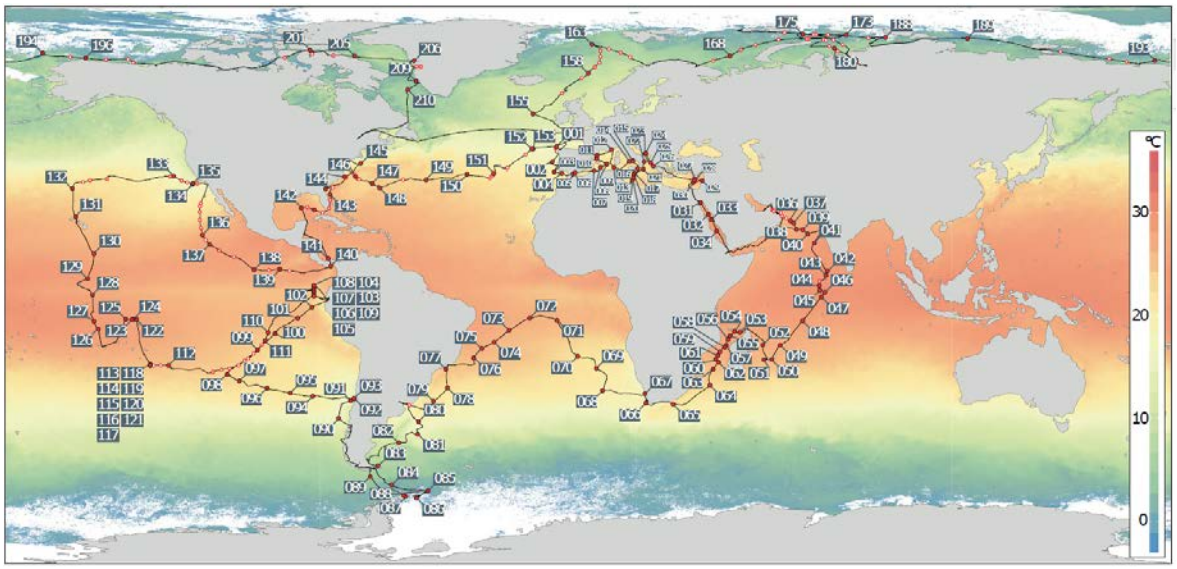

Figure 1.1: Cruise track of the Tara vessel from September 2009 to December 2013 and the location of 210 stations (extracted from Sunagawa et al., 2020).

The sampling strategy of *Tara* Oceans is exhaustively described in *Pesant et al. (2015)*. The choice of each station's placement was the result of the interpretation of in situ data, satellite and literature information collected to define oceanographic structures, weather and seasonal conditions occurring in a precise time and space, with the overall goal that each sample would capture a single picture of specific conditions such as mesoscale eddies, upwellings, anaerobic zones, main currents and acidic waters, biodiversity hotspots or climatically relevant locations. For each station, water samples were size-fractionated using filters of different diameter to optimize the concentration of organisms into 12 size fractions going from femto- to megaplankton, and from different depths including the subsurface layer, the deep-chlorophyll maximum (DCM), and the mesopelagic zone. Back in the laboratory, collected filters were used to extract nucleic acids that underwent high-throughput sequencing (HTS) to generate metabarcoding, metagenomic and metatranscriptomic data sets as well as to yield single-cell genomes. All the sequencing protocols used for *Tara* samples are reported in *Alberti et al. (2017)*.

The project applied the most advanced methodologies to sample and process data, describing plankton communities in their environment at an unprecedented

resolution and scale. In particular, molecular approaches used and developed in the context of *Tara* Oceans have been fundamental in defining plankton biogeography in absence of morphological features and gave a strong contribution to the discovery of new lineages and their distribution patterns. Cutting-edge methodologies like DNA metabarcoding, metagenomics, metatranscriptomics and single cell genomics have been used and implemented, leading to the construction of some of the richest databases of planktonic molecular data. Metagenome-Assembled Genomes (MAGs) have been recently reconstructed from *Tara* prokaryotic and eukaryotic metagenomes, providing new insights into planktonic diversity and evolution (*Royo-Llonch et al., 2020*; *Delmont et al., 2022*).

1.4.1.1 Metabarcoding

Communities of microorganisms can be studied through environmental metabarcoding, an amplicon sequencing approach where a fragment of a gene is amplified and massively sequenced from an environmental sample. The sequenced gene is usually a fragment of the small subunit of the rRNA gene, that contains several hypervariable regions flanked by relatively conserved ones and is therefore suitable for taxonomic assignments. Metabarcoding reads from *Tara* Oceans consisted of the V4-V5 regions of the 16S rDNA for prokaryotes, and the V4 and V9 regions of the 18S for eukaryotes (*Alberti et al., 2017*, *Ibarbalz et al., 2019*).

In 2015, the first large-scale metabarcoding survey based on V9-18S rDNA provided the identification of about 150,000 eukaryotic taxa in the global epipelagic ocean, out of which ~33% were previously unknown (de *Vargas et al., 2015*). High diversity emerged at all taxonomic levels, from phylum to genus, with a high biodiversity held by heterotrophic protists and probably explained by their biotic interactions rather than by competition for resources and space. This survey has become a reference source for phylogenetic and biogeographical studies at regional and global scales for planktonic communities and the dataset provided has also been used for developing and testing different bioinformatic tools

(*Callahan et al., 2017*; *Foster et al., 2017*; both reviewed in *Sunagawa et al., 2020*). Recently (*Ibarbalz et al., 2019*), the global V9 metabarcoding dataset was used together with the V4 information and implemented with data from *Tara* Oceans Polar Circle expedition to study one of the fundamental patterns in biogeography and macroecology: The Latitudinal Diversity Gradient (LDG), i.e., the monotonic poleward decline of local diversity for both terrestrial and aquatic organisms (*Hillebrand, 2004*). This work showed that most planktonic functional and taxonomic groups inhabiting the euphotic ocean exhibit a poleward decrease of biodiversity, while this is not true below the Deep Chlorophyll Maximum (DCM). Temperature has been identified as the best predictor of such patterns, raising important considerations on how global warming may increase plankton diversity especially at high latitudes. Metabarcoding was also used to study species distribution and diversity for specific groups: dinoflagellates showed regular patterns across the global surface ocean (*Le Bescot et al., 2016*), while epipelagic diatoms were made by few cosmopolitan OTUs with even distribution and high abundance, many of which unknown (*Malvyia et al., 2016*).

The richness of this dataset in terms of sampled ecosystems allowed assessments of patterns that were above the mere geographic ones, and that could be linked to specific environmental features; the biflagellated unicellular diplonemids, for example, although not showing any geographic structuring, exhibit a clear vertical stratification, with higher abundance and diversity in the deep mesopelagic ocean (*Flegontova et al., 2016*), while Collodaria (phylum: Radiolaria), whose communities are homogeneous in each biome, showed diversity patterns according to the trophic status of the oceanic provinces, with coastal waters being less diverse than the open oligotrophic ocean. Other than describing biogeographical patterns of the most abundant taxonomic groups, *Tara* Oceans metabarcoding data allowed to make a first evaluation of mixotrophs, a functionally and structurally heterogeneous group that is rapidly changing our view of the traditional plankton trophic web that assumes a separation of trophic processes between different

organisms (*Flynn et al., 2019*). Using *Tara* Oceans V9-18S metabarcoding region, *Faure et al. (2019)* have found that mixotrophy is a ubiquitous feature but there are differences depending on the functional type of the mixotroph.

1.4.1.2 Metagenomics

Metagenomics, like metabarcoding, is a DNA-based technique. The main difference between the two is that metagenomics relies on the sequencing of genomic fragments from the whole DNA and not only one gene, allowing the detection of thousands of genomes directly from their natural habitats. With the simultaneous improvement of high-throughput sequencing technologies and their progressive cost reduction, and with the development of more effective bioinformatic pipelines, metagenomics has revolutionized plankton research.

Metagenomic approaches require the comparison of sampled DNA, in the form of a set of reads, to reference genomes, in order to perform a verified taxonomic assignation; the main limit of this method is thus the lack of completeness of sequenced genomes, together with the lack of universal standards to build the genomic databases (*Loeffler et al., 2020*). Therefore, many metagenomic studies, including those on plankton biogeography using *Tara* Oceans data, rely on fragment recruitment, a process of aligning metagenomic sequencing reads to available reference genomes. In *2013, Swan et al.* studied the global distribution of surface ocean bacterioplankton using Single-Cell Amplified Genomes (SAGs) as reference genomes in metagenomic fragment recruitment, and demonstrated that bacterioplankton biogeography correlates with temperature and latitude and is not limited by dispersal, in agreement with the global ubiquity hypothesis (*Baas-Becking, 1934*, see section 1.2.3). In *2016, Vannier et al.* used one available genome and one de novo SAG of the Mamiellales genus *Bathycoccus* for metagenome fragment recruitment, highlighting their similarities (e.g., The two genomes share the same V9 region of 18S rRNA gene) and differences and showing how the two genomes rarely co-occur in the global epipelagic ocean but instead occupy distinct niches,

mainly shaped by depth. The non-random distribution of Mamiellales species has also been confirmed in a recent study by *Leconte et al. (2020)*, that highlighted the role of temperature as a main factor in shaping the global biogeography of this important order of green algae. Metagenomic reads have also been mapped against the Global Ocean Viromes dataset (*Gregory et al., 2019*), allowing for the detection of a viral biogeography and the identification of a biodiversity hotspot for viruses in the euphotic Arctic Ocean. More recently, the high uniqueness in polar biomes has been specifically assessed for Nucleocytoplasmic Large DNA Viruses (NCLDVs), that display a heterogeneous distribution across oceans and a vertical similarity in the water column (*Endo et al., 2020*). Finally, metagenomic data have been used to confirm the amphitropical distribution of the copepod *Oithona*, also allowing the study of its population structure in the Mediterranean Sea (*Madoui et al., 2017*).

1.4.1.3 Metatranscriptomics

Together with metagenomics, metatranscriptomic data was produced from each filter sampled during *Tara* Oceans expeditions. As it happens for the first, the latter starts with random extraction of nucleic acid from the sample, with the substantial difference that metatranscriptomics require mRNA instead of DNA extraction, thus targeting the genes that are actually expressed by an organism. Therefore, metatranscriptomic data allows for the observation of the global response of entire communities to environmental conditions.

Tara Oceans metagenomic and metatranscriptomic sequencing effort resulted so far into the production of two main ocean gene catalogs: (i) the Ocean Microbial Reference Gene Catalog (OM-RGC) and (ii) the Marine Atlas of *Tara* Oceans Unigenes (MATOU). The first is a collection of 40 million non-redundant genes from viruses, prokaryotes and picoeukaryotes smaller than 3 µm retrieved from public marine plankton metagenomes and reference genomes (*Sunagawa et al., 2015*). The MATOU is a catalogue of 116 million unigenes obtained from poly-A+ cDNA

sequencing of different filter size fractions ranging from 0.8 to 2000 µm (*Carradec et al., 2018*). These catalogues also include metagenomic information, which enables to relate both gene abundance and expression and to study them in relation to environmental parameters for a better understanding of communities' composition and functions (*Salazar et al., 2019*; *Vorobev et al., 2020*). These newly established resources are being periodically updated as genomic and transcriptomic databases are publicly released, and have already significantly enhanced our knowledge of the planktonic gene repertoire, allowing studies on biogeography of genes and functions across oceans at a global scale. For instance, recent research on prokaryotes (*Salazar et al., 2019*) correlated expression of genes involved in carbon sequestration and photosynthesis with both water depth and geographical patterns. Understanding the regulation of gene expression is more challenging when looking at eukaryotes; however, several studies based on *Tara* Oceans metatranscriptomics shed light on eukaryotic plankton response to environmental factors addressing the expression of key ecological genes involved in the metabolism of important micronutrients, like the case of iron (*Caputi et al., 2019*) or macronutrients like nitrate (*Busseni et al., 2019*).

The whole work of the present thesis has been developed exploiting this unprecedented amount of molecular and environmental data provided by the *Tara* Oceans expedition.

1.5 Thesis overview

Starting from the conceptual challenges and the topics treated in the present chapter, the general aim of my thesis has been the investigation of diatom diversity, biogeography and functioning at global and regional scale from community to species level through the application and development of different methodological frameworks. This study was mainly based on data sets provided by *Tara* Oceans

and *Tara* Oceans Polar Circle expeditions, including the metabarcoding, metagenomics and metatranscriptomic datasets, as well as the environmental metadata.

The following two chapters provide a community-level description of diatoms; in particular, Chapter II describes a first assessment of diatom communities by measures of alpha-diversity indices and their relationships, together with the analysis of species abundance distributions (SADs) patterns from samples collected at global scale. In this chapter I applied the main principles of community ecology following a macroecological approach to reveal emergent patterns in diatom communities' structuring. This kind of study is normally intended for metabarcoding data, thus targeting Operational Taxonomic Units (OTUs); I extended this exercise to metagenomics and metatranscriptomics, providing a first statistical description of the structure and properties of the different types of high throughput sequencing data, and discussing its implications.

To investigate the effects of environmental variation on diatom communities' composition and gene expression, in Chapter III I focused on the case study of the North-Atlantic – Arctic Ocean transect, by analysing how diatom communities change along the current that represents the main inflow in the Arctic Ocean, a region particularly vulnerable to climate change. Here I integrated environmental in situ data collected during *Tara* Oceans expedition with other two sources, i.e., the World Ocean Atlas (*Boyer et al., 2018*) and PISCES model (*Aumont et al., 2015*), also providing an overview in similarity and differences among the three. In this context, metabarcoding information was used to study compositional diversity of stations; by integrating a phylogeny-based approach, I also exploited metatranscriptomic data to study the variation in abundance and expression of diatom genes involved in key metabolic processes like iron uptake, transport, and storage.

After the study of diatoms from a comprehensive and community-oriented perspective, the last two chapters provide a deep investigation of the ecologically relevant diatom genus *Pseudo-nitzschia*. In Chapter IV, the construction of a metabarcoding dataset based on a curated reference source allowed the description of the global distribution of *Pseudo-nitzschia* species with an unprecedented resolution. This biogeography was then explored from an ecological point of view with the integration of abiotic and biotic factors. In particular, by incorporating metabarcoding information from the prokaryotic world, I could build co-occurrence networks for the investigation of *Pseudo-nitzschia* -bacteria in silico interactions.

Finally, in Chapter V I performed a functional study focused on genes whose expression is induced during sexual reproduction in *Pseudo-nitzschia* and other diatoms. I explored the abundance and distribution of these genes in the context of *Tara* Oceans to test their role as potential molecular markers to detect the elusive event of sexual reproduction at sea. To conclude, in Chapter VI I discussed the main outcomes of the former chapters, the problems encountered and the overall scientific contribution provided by this work.

Chapter II

Diatom communities through a macroecological approach

2.1 Introduction

Macroecology is a sub discipline of ecology that aims to find and predict universal ecological patterns and to describe their origin and maintenance across taxa and ecosystems and spatiotemporal scales.

Besides its large-scale nature and a strong emphasis on observational data, the feature that more than others describes the essence of macroecology is its holistic approach, based on the assumption that properties at large scales are in most cases unpredictable from the investigation of what happens at smaller scales, but rather can only be understood when studying the system as a whole (*Brown, 1995*; *McGill, 2019*). This view is opposite to reductionism, a common practice in ecology that breaks a system up into its constituent parts and uses increasingly higher-resolution tools to study the details of the behaviour of the isolated components and their interactions in order to understand the whole. The key feature of macroecology is thus the ability to stand back, take a distant view from the studied phenomenon and look at the bigger picture to reveal emergent patterns and processes (*McGill, 2019*). According to this view, emergent patterns are ruled by universal law-like mechanisms intrinsic to life itself that reflect the geometric, physical, chemical, and thermodynamic principles that control the living of biological entities across different conditions (*Marquet et al., 2007*). One example is

the Taylor's law (*Taylor, 1961*), that describes the scaling relationship between the mean density and variance of a population as a power law. This fundamental pattern has been confirmed in ecology for many taxa and applies to the description of species variations both in space and time (*Xiao et al., 2015*, and references therein); moreover, Taylor's law was recently observed for transcribed genes obtained through single-cell RNA sequencing (*Lazzardi et al., 2021*), a dataset biologically different from what is obtained through DNA-based metabarcoding and metagenomics, that represents not only the presence or abundance of organisms in a community, but reflects their biological and metabolic activity. Another widely observed macroecological pattern is the positive power-law relationship between species and area (SAR; *Arrhenius, 1920, 1921*; *Williams, 1943*), that links the richness (i.e. the number of different species or taxa) to the size of the occupied area. A similar widely observed relation involves species abundance with their occupancy (AOR): species that are very abundant within one site are also likely to occupy many sites, while those that are locally rare tend to not be detected elsewhere (*Gaston et al., 2000*). Finally, species-abundance and species-rank-abundance distributions (SADs and RADs, respectively; see section 1.2.2), that display the abundance of each species within a community, show a highly conserved pattern that holds across different taxa and ecosystems: the majority of communities are uneven, with few very abundant species and many rare species (*Fisher et al., 1943*; *Preston, 1948, 1962*; *Whittaker, 1965*; *Brown, 1995*; *Gaston and Blackburn, 2000*). The debate of what mechanisms generate those patterns is still challenging ecologists, with evidences from several contrasting models, based on opposite views from neutral and niche theories, as equally well-fitting SAD patterns (*Etienne and Olff, 2005*; *Chisholm and Pacala, 2010*; *May et al., 2015*).

The above-mentioned laws and patterns have been originally discovered and successively confirmed mainly for terrestrial macroorganisms; however, the development of high-throughput DNA sequencing methods and the following accumulation of microbial datasets, together with the advent of technologies to

easily manipulate and model microorganisms allowed extending the main macroecological concepts to the world of microorganisms, leading to the birth of microbial macroecology. This discipline assumes that mechanisms behind large-scale patterns in microorganisms are similar to those applied for macroorganisms: dispersal limitation, neutral processes, environmental filtering, evolutionary and plastic responses, and local extinctions (*Xu et al., 2020*). However, this assumption has been much debated; on one side, the difficulty in identifying or defining microbial individuals due to their frequent modular organization, and to apply the species concept to organisms characterized by a preponderance of parasexuality (*Shade et al., 2018*) support the idea that a classic macroecological framework cannot be merely transposed to microbial communities as it is. On the other side, the taxonomic assignation of plants and animals to species is not always straightforward or precise, and some of the challenges are being encountered for macroorganisms too, especially when analysing insects. An overall consideration is that, despite the differences, from a more technical point of view both micro- and macro-organism's community data are traditionally recorded in a matrix where each column corresponds to the presence or abundance of species in the sampling unit (descriptor) and each sampling site or observation (object) occupies one row (*Legendre and Legendre, 2012*). From this table, regardless of the size of the organisms, macroecological patterns can be investigated, and their variation over space or time can be addressed.

For instance, a recent study (*Shade et al., 2018*) showed that the abundance-occupancy relationship (AOR) was maintained both for microbiota sampled from human belly buttons and for birds observed in the Czech Republic. Moreover, the species-area relationship, also called taxa–area or gene–area relationship (indicated respectively as SAR, TAR and GAR), whose application to microbial communities has important implications for global biodiversity (*Meyer et al., 2018*), has been confirmed for microorganisms. In particular, this relationship has been observed for both prokaryotes (*Fierer and Jackson, 2006*; *Horner-Devine et al., 2004*;

Zhou et al., 2008) and eukaryotes (*Green et al., 2004*), through morphological and metabarcoding approaches and also through microarray-obtained genes (*Zhou et al., 2008*). Furthermore, a recent study on gut microbiome community composition associated with reptiles, birds and mammals (*Godon et al., 2016*) showed a consistent increase in microbial diversity with increasing animal masses, a pattern consistent with the area–species theory (*Kieft, 2017*). The comparison of macroecological patterns in both macro- and micro-organisms also revealed similar species abundance distribution (SAD) curves (*Shade et al., 2018*), with microbial SADs showing the above discussed general hollow shape that reflects the uneven structure of communities, where dominant species are a few and rare species are numerous. However, notable differences related to higher rarity in the microbial realm were detected. Rarity is a community feature related to the skewness, a statistical property often used to describe SAD curves (*Kondratyeva et al., 2019*): the skewness metric quantifies the asymmetry of a distribution, and microbial SADs invariably display a strongly right-skewed pattern. Rare species are known to be an incredibly large quota of microbial communities, although they could also be represented by dormant inactive cells. The use of both genomic and transcriptomic data has been recently suggested as an approach to help to solve this problem, by discovering if a rare taxon is active or not in the community (*Jia et al., 2018*). Besides that, the debate of whether rare species represent a reservoir of community diversity essential to effectively respond to environmental perturbations (*Caron and Countway, 2009*; *Lynch and Neufeld, 2015*) or simply arise as an artefact of PCR and sequencing methods (*Brown et al., 2015*; *Elbrecht et al., 2018*) has yet to be resolved. A recent large-scale study on both macro- and micro-organisms' communities using high-throughput DNA sequence data (*Locey and Lennon, 2016*) confirmed once more the fundamental contribution of the rare biosphere to the SAD and found that, for a given community size, the quantity of rare species is always higher for communities of microorganisms than for macroorganisms. Moreover, this study found consistent power-law relationships between different aspects of species diversity, including rarity, evenness,

dominance and richness, and the total individual abundance (N) within ecological communities and microbiomes. These relationships hint at the existence of a unified scaling law that applies to both macroscopic and microscopic worlds. A step further toward the formalization of microbial macroecology has been recently made by *Grilli (2020)*, with the identification of three macroecological laws able to explain temporal and intercommunal species abundance fluctuations. These laws tackle some of the aspects I described before: the first law clarifies the relation between abundance and occupancy, the second validates the Taylor's law for microbial communities, and the third suggests that community variability is not explained by neutral processes, but it is rather mainly driven by environmental fluctuations.

Some general macroecological patterns and relationships, traditionally studied in terrestrial organisms, also hold in the sea, where peculiar physical forcing like horizontal advection and vertical mixing by ocean currents influence variation in species ranges, abundance and diversity (*Wieters, 2001*). For instance, the species–area relationship (SAR) and the abundance–occupancy relationship (AOR), appear to occur across both marine and terrestrial systems (*Drakare et al., 2006*; *Webb et al., 2011*). Likewise, similar patterns of species-abundance distributions (SADs) emerged from a study that compared marine benthos and fish with terrestrial micro fauna and ant communities, with all the assemblages dominated by rare species (*Gray et al., 2006*). For plankton, drifters immersed in a global-scale fluid environment, the role of the so-called seascape, i.e., the combination of physical, chemical, geological, and biological factors, is especially important in shaping the complex macroecological patterns. In general, large-scale dispersal processes and environmental filtering have been shown to shape the spatiotemporal structuring of planktonic bacteria (*Hellweger et al., 2014*) and protists (*Foissner, 2006*; *de Vargas et al., 2015*), including diatoms (*Casteleyn et al., 2010*). Size, ecology and geographical distribution of planktonic organisms have been linked in a recent study (*Sommeria-Klein et al., 2021*) that showed how larger heterotrophic organisms have more

restricted ranges than small-sized autotrophs, which, in turn, are globally structured by latitude but also strongly influenced by local conditions. Macroecological studies on plankton also investigated the large-scale patterns of the rare biosphere components (*Ser-Giacomi et al., 2018*), whose SADs showed a power-law decay constant across geographical regions, pointing to the major role of dispersal processes in governing the biogeography of rare biosphere. Rare taxa were also recently accounted for in a global-scale investigation of spatial patterns of diatom richness (*Busseni et al., 2020*). Finally, a study on freshwater diatoms (*Passy, 2016*) highlighted how macroecological patterns in diatom species-abundance distributions, as well as in richness, are jointly controlled by local environment and large-scale climatic factors, emphasising the need for a stronger focus on the role of climate in driving diatom distributions (*Soininen and Teittinen, 2019*).

Altogether, these findings show how the exponential increase in available microbial data and the rapidly growing awareness of how ocean microbial communities are structured is leading to the emergence of unifying principles that bridge traditional macroecology with microbial macroecology; this gradual and progressive incorporation of microorganisms into macroecological theory is likely to expand our knowledge of global ecological patterns, allowing us to better understand the fundamental processes responsible of abundance and diversity patterns across all forms of life, and to improve the predictions of organisms' responses and feedbacks to global change (*Shade et al., 2018*; *Xu et al., 2020*). This understanding cannot be separated by the integration of functional information, its variation across space and time as well as its separation from the pure taxonomical information (e.g., *Louca et al., 2016*). Moreover, the integration of transcriptomic data in this framework has the almost unexplored potential to provide precious insight to the formulation, the confutation or the confirmation of universal laws in ecology. In this context, the present chapter aims to provide a general overview of the structural properties of diatoms' communities in the global surface ocean. I will explore metabarcoding, metagenomic and metatranscriptomic data from the *Tara*

Oceans and *Tara* Oceans Polar Circle expeditions in order to provide a first description of the macroecological patterns deriving from the analysis of different high-throughput sequencing data. In particular, I will both visually compare the distribution of the data types and synthesize them through summary statistics, and I will focus on diatom biodiversity studying how diatom abundance (N) constraints richness, dominance, rarity and evenness among communities, and evaluate if these diversity metrics scale with N.

2.2 Methods

2.2.1 Data

Metabarcoding, metagenomic and metatranscriptomic data collected across *Tara* Oceans and *Tara* Oceans Polar Circle expedition were exploited in the present chapter. In particular, the metabarcoding information was used in the form of Operational Taxonomic Units (OTUs) obtained through swarm clustering (*Mahé et al., 2014*) of sequencing reads representing the V9 region of the 18S rDNA gene. Unigenes belonging to the MATOU (Marine Atlas of *Tara* Oceans Unigenes; *Carradec et al., 2018*) were analysed with respect to their abundance and transcription values in metagenomic and metatranscriptomic data, respectively. In particular, I used a geographically extended version of the published MATOU, namely MATOU-v2, that included the meta-omic information from the Arctic Ocean, collected during the *Tara* Oceans Polar Circle expedition. The biological units used, being them OTUs or unigenes, were all taxonomically assigned to diatoms (Bacillariophyta) and belonged to size-fractionated surface samples corresponding to the size class 20-180 µm.

Using information on geographical coordinates, together with depth and collection date, environmental data were extracted from the World Ocean Atlas 2018

(WOA18; *Boyer et al., 2018*) and from PISCES model (*Aumont et al., 2015*) for each selected station. Temperature, salinity, as well as nutrient concentration of nitrate, phosphate and silicate were extracted from WOA18, while concentration of micronutrients such as cobalt, copper, iron and manganese were obtained from PISCES. Data extraction from WOA18 was performed by Paul Frémont (Genoscope, France), while the information from PISCES model was provided by Dr. Alessandro Tagliabue (University of Liverpool, UK).

2.2.2 Filtering

Metabarcoding, metagenomic and metatranscriptomic data (respectively indicated as metaB, metaG and metaT) were filtered to remove the most locally rare units. In particular, I filtered out from each sampling station every OTU or unigene that occurred with an absolute abundance, expressed as read counts, less than 4. Moreover, metaG and metaT were filtered in order to compare the same set of genes in each sample. In theory, every unigene detected in metaT would be matched by its occurrence in the corresponding metaG sample, as mRNA is the result of transcription of the genomic information contained in the DNA. However, due to technical reasons (large genome size, under sampling, low coverage), this was not always the case. On the other side, the presence of a unigene in metaG and not in the corresponding metaT sample would mean that the gene is held by the organism but not actively expressed at that location and time. One aim of the present chapter is to evaluate the difference between the two meta-omics-derived information in the context of the macroecological patterns investigated; therefore, I decided to keep only unigenes occurring in both metaG and metaT at each station. This choice inevitably caused a loss of information and statistical robustness, but allowed a coherent comparison between the meta-omic DNA and mRNA information at the single station level.

The analysis presented in this chapter were also performed using data obtained through a more relaxed filtering approach, that kept the same threshold on

abundance and retained the same set of genes in metaG and metaT on a global scale, thus not accounting for possible discrepancies at local scale. Moreover, a more stringently filtered data set was obtained by increasing the value of the abundance threshold from 3 to 10 and considering only genes occurring in both metaG and metaT in each station.

The final dataset used for the analysis was composed by 5308 diatom OTUs and 2,987,248 diatom unigenes, distributed across 77 *Tara* Oceans surface sampling stations belonging to the 20-180 µm size fraction, corresponding to micro-plankton.

2.2.3 Diatom richness and abundance at global scale.

Diatom richness, expressed as the number of distinct OTUs and unigenes, as well as diatom abundance, indicated by the absolute number of read counts, were calculated for each station and plotted on a geographical map. Moreover, the relative fraction of diatom OTUs and unigenes out of the total abundance of OTUs and unigenes for all other planktonic taxa was calculated for each station of interest. This analysis provided an overview of the importance of diatom abundance in term of taxa and genes present (metaB and metaG) and the relative contribution of diatom metabolic activity (metaT) in the global surface ocean. Diatom richness from metaB, metaG and metaT was correlated with environmental data. The environmental parameters used for the correlations are listed in table 2.1.

Table 2.1. Environmental descriptors selected for this study. Values extracted from WOA18 and PISCES are monthly mean surface (5 m depth) values.

Parameter	Symbol	Unit	Source
Temperature	Temp	°C	WOA18
Salinity	Sal	PSU	WOA18
Nitrate	NO_3^-	µmol/L	WOA18
Phosphate	PO_4^{3-}	µmol/L	WOA18

Silicate	$Si(OH)_4$	µmol/L	WOA18
cobalt	dCo	picomol/L	PISCES2
copper	dCu	nmol/L	PISCES2
iron	dFe	nmol/L	PISCES2
manganese	dMn	nmol/L	PISCES2
zinc	dZn	nmol/L	PISCES2

2.2.4 Species Abundance Distribution

For each sampling site, diatom OTUs and unigenes were explored using species abundance distribution (SAD; *Preston, 1948*) and rank-abundance distribution (RAD; *Whittaker, 1965*) curves, as they are the two main approaches used in ecology to display the structure of a community. Through SAD visualization, abundance intervals are built, each interval being twice as wide as the preceding one (but presented as a standard histogram with bars of equal breadth), and the number of species within each interval is graphed. This process is equivalent to first log_2 transforming the data and then constructing a frequency histogram. In RAD plots the abundance of each species within the community is ordered from the most abundant to the rarest and the proportion of the contribution of each species to the total assemblage is plotted against the rank order of that species. The shape of the RAD curve summarizes the abundance of each species in a community, thus providing simultaneous information on richness and evenness; I presented RAD curves with a logarithmically transformed abundance axis, which emphasizes the part of the distribution describing the least abundant species.

2.2.5 Statistical descriptors

In order to compare values from the three datasets explored here, I converted the metabarcoding, metagenomic and metatranscriptomic read counts into their relative abundance values, through the total sum scaling transformation, which divides the read counts by the total count in each sample.

Once the three datasets were made suitable for comparison, I calculated four statistical descriptors, known as "moments" of a distribution, to study diatom OTUs and unigenes across *Tara* Oceans sampling stations: mean, variance, skewness and kurtosis (Table 2.2). The mean represents the central value, while the variance measures how far the data values are spread out from their central value. In order to meaningfully compare different distributions, I used the ratio of the variance to the mean. A further characterization of the data included skewness and kurtosis. Skewness is a measure of the asymmetry of a distribution, while kurtosis measures whether the data are heavy-tailed or light-tailed relative to a normal distribution: low kurtosis indicates light tails, or lack of outliers, while high kurtosis indicates the presence of outliers at the extreme values of the distribution. As kurtosis and skewness effectively describe the shape of distributions, they are good parameter choices for clustering distributions into profiles according to shape properties. All statistics were calculated using R 4.0.3 (*R Core Team, 2020*) and the *e1071* package (*Meyer et al., 2021*).

Table 2.2. Formulas used to calculate the moments of distributions.

Metric	Formula
Mean	$\sum_{i=1}^{n} x_i/n$
Variance	$\sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)$
Skewness	$b_1 = \frac{m_3}{s^3} = \left(\frac{n-1}{n}\right)^{3/2} \frac{m_3}{m_2^{3/2}}$
Kurtosis	$b_2 = \frac{m_4}{s^4} - 3 = \left(\frac{n-1}{n}\right)^{2} \frac{m_4}{m_2^2} - 3$

s is the standard deviation
n is the sample size

$$m_r = \frac{1}{n} \sum (x_i = \bar{x})^r$$

2.2.6 Scaling of biodiversity indices with total abundance

For each sampling site biodiversity was estimated using the three data sets through four indices: richness, dominance, rarity and evenness.

Richness

Richness (S; Table 2.3) counts the number of different species present in a sample. Here, I applied this index to measure the number of different diatom OTUs or unigenes in a sampling station, as richness represents simultaneously the simplest and most widely used metric to estimate biodiversity (*Whittaker, 1972*; *Magurran, 2013*; see section 1.2.2).

Dominance

Dominance calculates the importance of the species that contributes the most to total abundance or biomass within an assemblage and can be expressed in its absolute or relative form, depending on whether the computed index is divided or not by total abundance. Here I estimated absolute dominance as the number of reads of the most abundant OTU or unigene in each of the three datasets using Berger-Parker index (BP; *Berger and Parker, 1970*; Table 2.3).

Rarity

Rarity indices measure the proportion or the quantity of the least abundant species within a community. Distributions of species abundance commonly display a strongly right-skewed shape — that is, a positive skewness — and rarity is usually calculated on a $\log_{10}$ scale. Therefore, to prevent numerical inconsistencies, I first took the absolute value of the skewness, then added 1, and finally calculated its logarithm — applying what is commonly called log-modulus transformation (as in *Locey and Lennon, 2016*; Table 2.3).

Evenness

Evenness emphasizes the similarity in numbers of each species in an environment: low values reflect the presence of one or few dominant species, while high values indicate that different species have similar numbers of individuals and the distribution of abundance is uniform. Evenness can be measured in many different ways (*Alatalo, 1981*; *Pierson, 2020*) and it is not an independent measure; it can be derived from compound diversity measures that inherently contain richness and evenness components, such as Shannon's diversity (H'; *Shannon, 1948*), Simpson's diversity and Simpson's dominance (respectively D_1 and D_2; *Simpson, 1949*). Here I used Simpson's evenness index (E; *Simpson, 1949*; Table 2.3), obtained by dividing Simpson's dominance (D_2) by richness (S); this metrics is mathematically independent from Simpson's diversity (*Smith and Wilson, 1996*) and therefore represents a useful measure of evenness in many contexts (*Morris et al., 2014*).

Once defined the indices, I tested whether the total abundance of diatom OTUs or unigenes within a sampling station was a constraint on biodiversity as measured through its richness, dominance, rarity and evenness components. Assuming that N is a variable across which the biodiversity components scale (*Lennon and Locey, 2016*), I examined these relationships using a simple linear regression on log-transformed axes. The scaling law, also referred to as power law, is expressed by the formula $y \sim x^z$. This relationship becomes linear when both the independent and dependent variables are log-transformed: $y \sim x^z$ is in fact equivalent to $\log(y) \sim z\log(x)$, where z is the scaling exponent of the power law relationship. The representation using linear regression on log-transformed axes thus allowed to quantify the rate at which the scaling of biodiversity indices with N occurred. The coefficient of determination (R^2) was also estimated for each linear regression to measure the goodness-of-fit. All indices were calculated using R 4.0.3. (*R Core Team, 2020*) and the *vegan* package (*Oksanen et al., 2020*).

Table 2.3. Formulas used to calculate diversity measures studied.

Metric	Formula
Richness (S)	Number of OTUs or unigenes
Total abundance (N)	Number of reads
Shannon's diversity (H')	$-\sum_{i=1}^{S} p_i \ln p_i$
Simpson's diversity (D_1)	$1\text{-}\sum_{i=1}^{S} p_i^2$
Simpson's dominance (D_2)	$1/\sum_{i=1}^{s} p_i^2$
Berger–Parker dominance (BP)	p_{max}
Simpson's evenness (E)	D_2/S
Rarity	| Skewness | + 1

p_i - is the proportion of reads belonging to OTU or unigene i;
p_{max} is the proportion of reads belonging to the most abundant OTU or unigene.
$\bar{Y}$ is the mean
s - is the standard deviation
N - is the number of data points.

2.3 Results and Discussion

2.3.1 Filtering

The filtering step strongly influenced the number of OTUs and unigenes (Table 2.4). For instance, removing from each station OTUs with less than 4 reads led to a 59% reduction of the number of OTUs, that went from 3977 to 1616. This percentage of loss is similar to what was found for metagenomic data. A different pattern has been found for metaT, where the number of unigenes went from almost three million to a little more than 1 million, thus reducing the number of genes of 72%. It must be noted that the majority of this loss of genes in the metaT was not due to the imposed abundance threshold, but mainly to the second filtering procedure

applied to metaG and metaT, i.e., the selection of only genes that occurred in both omic data for each station.

Filtering data is a common practice in metabarcoding and meta-omic studies, that allows to remove further artefacts and errors not detected in previous steps of data processing. However, too stringent thresholds risk to filter out informative data on rare species, and empirical thresholds are being tested (*Faust et al., 2015*; *Burki et al., 2021*). The rationale behind the filtering procedure of the present chapter was, at first place, to make different data sets homogeneous for the comparison. The abundance threshold selected was not too stringent, but still allowed discarding the very locally rare OTUs and genes without removing them at global scale. The use of the same set of genes for metaG and metaT was necessary to meaningfully compare signals from the functional potential and the expressed activity of diatom communities (see section 2.2.2)

Table 2.4. Total Richness for metabarcoding, metagenomic and metatranscriptomic data before and after the applied filtering procedure.

Data	Unit	Unfiltered	Filtered
metaB	OTUs	3.977	1.616
metaG	unigenes	2.987.248	1.278.194
metaT	unigenes	4.491.688	1.278.194

2.3.2 Diatom richness and abundance at global scale

Richness, expressed as the number of different OTUs or unigenes in a sample, was plotted in logarithmic scale in Fig. 2.1. The Figure shows an overall agreement between patterns obtained with the three data sets: diatoms were abundant and displayed a high diversity at high latitudes, and especially in polar and subpolar regions of both hemispheres.

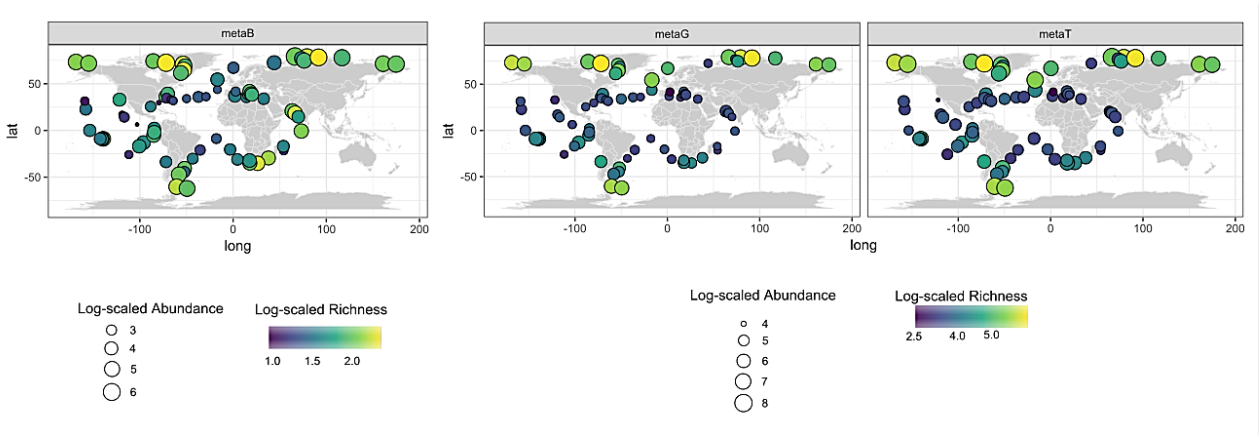

Figure 2.1. Global-scale diatom richness and abundance using metabarcoding (metaB), metagenomic (metaG) and metatranscriptomic (metaT) data.

Results in Fig. 2.2. show the relative contribution of diatoms to the whole sampled microplanktonic community, and confirm diatoms as highly abundant in cold and polar waters, where it is known they constitute the most important class of eukaryotic plankton (*Smetacek, 1999*). Diatoms are in fact well adapted to the main environmental stressors in polar seas: constant low and freezing temperatures, physical disturbances from sea ice, and extreme seasonality (*Cota, 1985*).

Although concordant in displaying the high polar abundance of diatoms, the three data sets showed some differences; in particular, metabarcoding data (Fig. 2.2.A) showed a high contribution of diatoms to the whole eukaryotic planktonic communities in some sampling stations of the Indian and South Pacific Ocean, which was not reflected in metagenomic or metatranscriptomic sets at these stations.

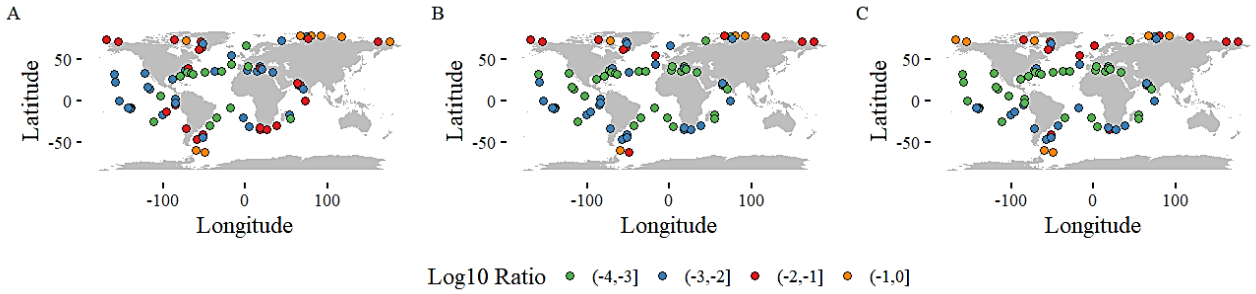

Figure 2.2. Log10 ratio of the relative contribution of diatom OTUs and unigenes to the whole micro-planktonic communities in the global surface ocean using A) metabarcoding, B) metagenomic and C) metatranscriptomic data.

Spearman's pairwise correlations (Fig. 2.3) were calculated between each pair of environmental parameters as well as for richness values in metabarcoding and meta-omic data (metaG and metaT have the same richness by construction). Richness values of metaB and metaGT were highly correlated. Moreover, they both showed strong negative correlations with temperature and salinity, in agreement with the high richness values of diatoms in the cold and low-salinity waters flowing across polar oceans. Richness patterns of metaG and metaT also showed strong positive correlations with nitrate and phosphate, two indispensable macronutrients for diatom metabolism.

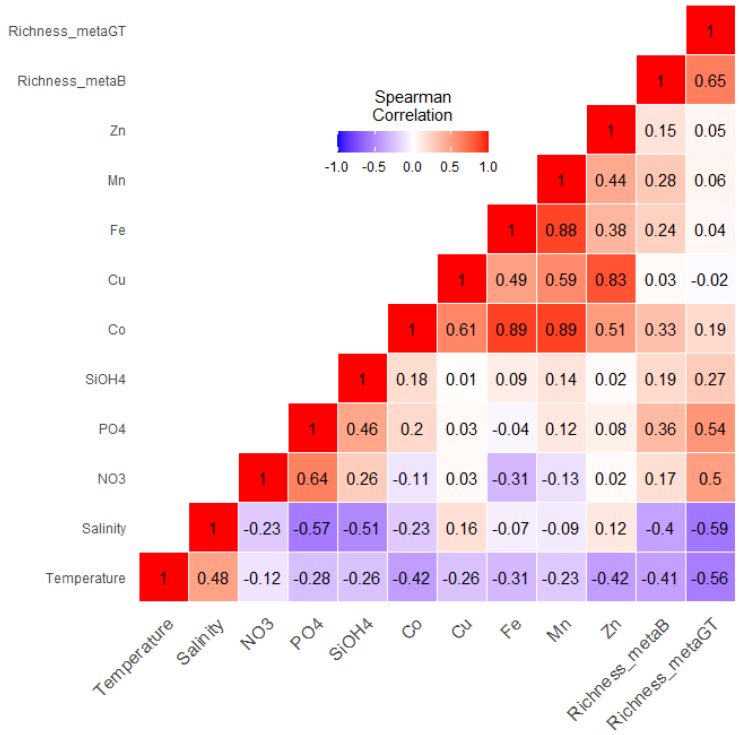

Figure 2.3. Heatmap showing pairwise Spearman correlation values between pairs of environmental variables and richness measures calculated for metabarcoding (metaB), as well as for metagenomic and metatranscriptomic data (collectively indicated as metaGT).

2.3.3 Species abundance distribution

Species Abundance Distribution (SAD) plots for one station, i.e., station 206, are shown in Fig. 2.4. The three plots show a right-skewed pattern, indicating that communities contained many rare OTUs and unigenes and only a few common OTUs and unigenes. SAD of genes occurring in metaT showed a different pattern, less skewed and more symmetrical, i.e., closer to a normal distribution.

Rank species Abundance Distribution (RAD) plots (Fig. 2.5) allowed a simultaneous visualization of how OTUs and genes were distributed in each sample; here each curve corresponds to a single station and the corresponding colour reflects the absolute richness of that station, calculated as the total number of different OTUs or unigenes. RADs showed a long right tail of rare entities for all three data sets. Metabarcoding data showed a higher variability of OTU distribution among stations compared to the one of unigenes found in metagenomic and metatranscriptomic

data, which showed a more compact structure. Results of metabarcoding data overall showed a strong regularity across stations, in agreement with what has been already observed for non-dominant taxa of plankton communities in a global-scale study (*Ser-Giacomi et al., 2018*). Although more compact, the behaviour of metagenomic data was similar to the one observed for diatom OTUs. Both metagenomics and metatranscriptomic RADs strongly resembled a power-law distribution, followed by an exponential tail. The same pattern was recently observed for transcriptomic data obtained through single-cell RNA sequencing data of mouse organs, where the shape of the distribution remained constant also when comparing datasets obtained through different laboratory procedures (*Lazzardi et al., 2021*). However, the shape of RADs for metaT showed an interesting relationship with richness. Two different main regimes could be detected, one held by the low-richness stations, and one characteristic of high-richness stations. This observation has never been found in metatranscriptomic data, and a step further in the generalization of the pattern I observed in the RAD analysis is currently being performed by Emanuele Pigani (Stazione Zoologica Anton Dohrn, Italy).

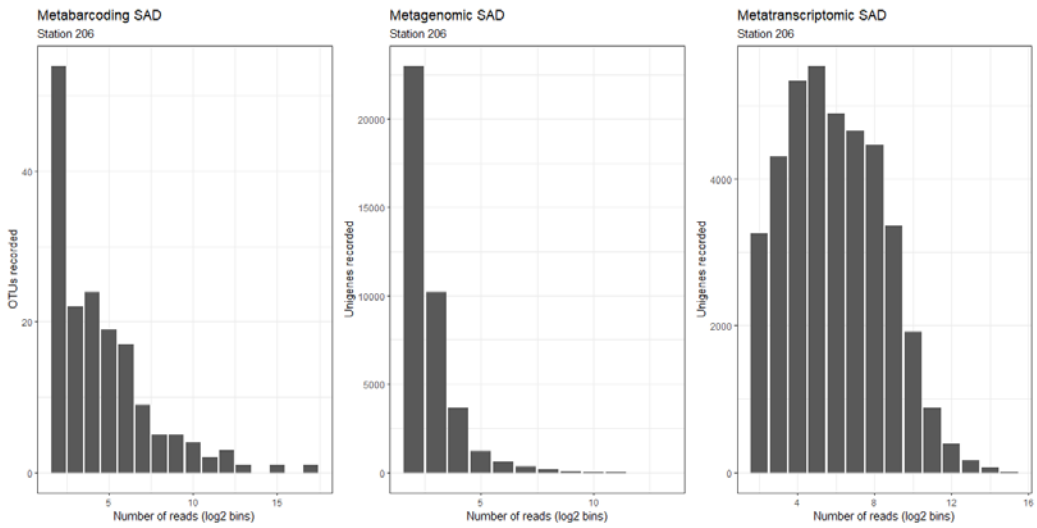

Figure 2.4. Species Abundance Distribution for diatom communities detected using metabarcoding, metagenomic and metatranscriptomic data.

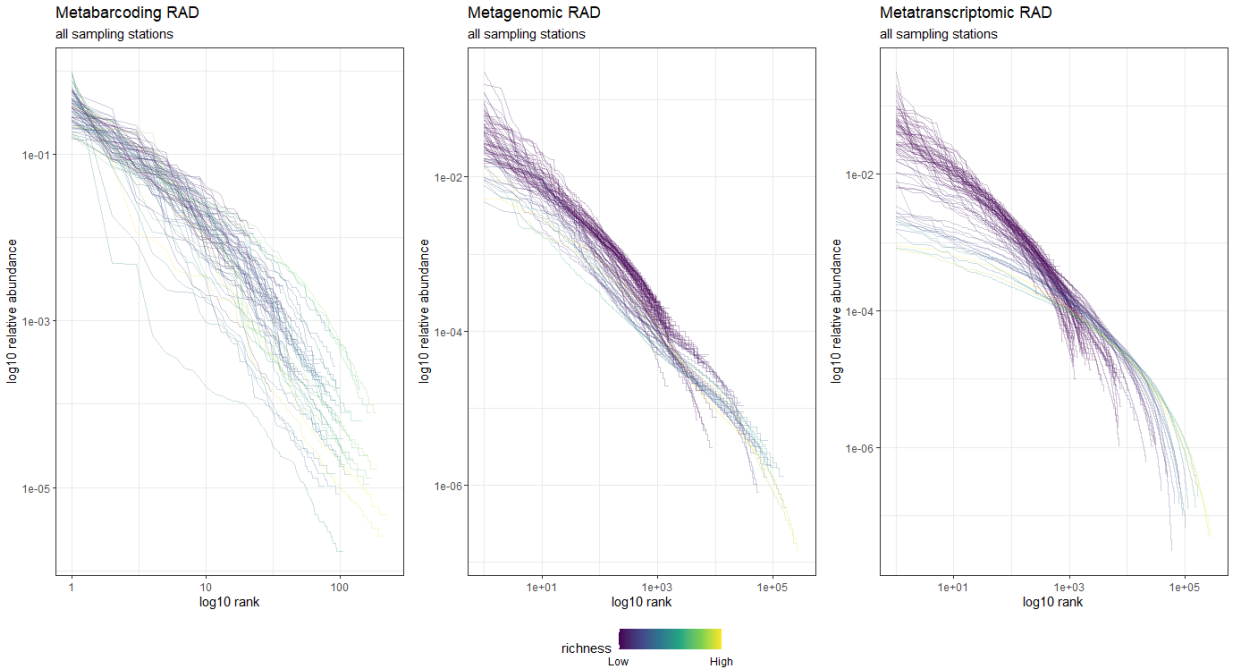

Figure 2.5. Rank Abundance Distribution for diatom communities detected using metabarcoding, metagenomic and metatranscriptomic data.

2.3.4 Statistical descriptors

In order to compare the distributions of metabarcoding, metagenomic and metatranscriptomic data sets, I synthesized them through the summary statistics known as moments of distributions, i.e., mean, variance, skewness, and kurtosis. During this phase, I compared each data type across stations while also comparing their global distribution.

Results of the relationship between mean and variance (Fig. 2.6.) confirmed the macroecological law known as Taylor's law, showing how variance of OTUs and unigenes scaled as a power-law function of the mean OTUs and unigene density. This pattern has been observed at different scales of biological organizations, from populations to single cells, including examples from both the prokaryotic and eukaryotic worlds (*Ramsayer et al., 2012*; *Azevedo et al., 2001*). Recently, a Taylor's law with a constant exponent equal to 2 has been reported for 7 different microbial community data sets analysed through metagenomics (*Grilli, 2020*). I found the same constant coefficient of variation of abundance fluctuations with respect to

mean in transcriptomic data; moreover, metagenomics data also scaled approximately quadratically with the mean (z = 1.4). The same did not occur for metabarcoding (z = 0.8). Furthermore, I found an inverse relationship of mean and variance with richness: rich stations are likely owing the longest tails of numerous rare species occurring with a very low abundance that pull down the mean value of the station and its variance. When looking at the three datasets together (Fig. 2.6D), a clear overlap of metagenomic and metatranscriptomic data was shown, while the metabarcoding mean and variance were shifted to values of three orders of magnitude higher.

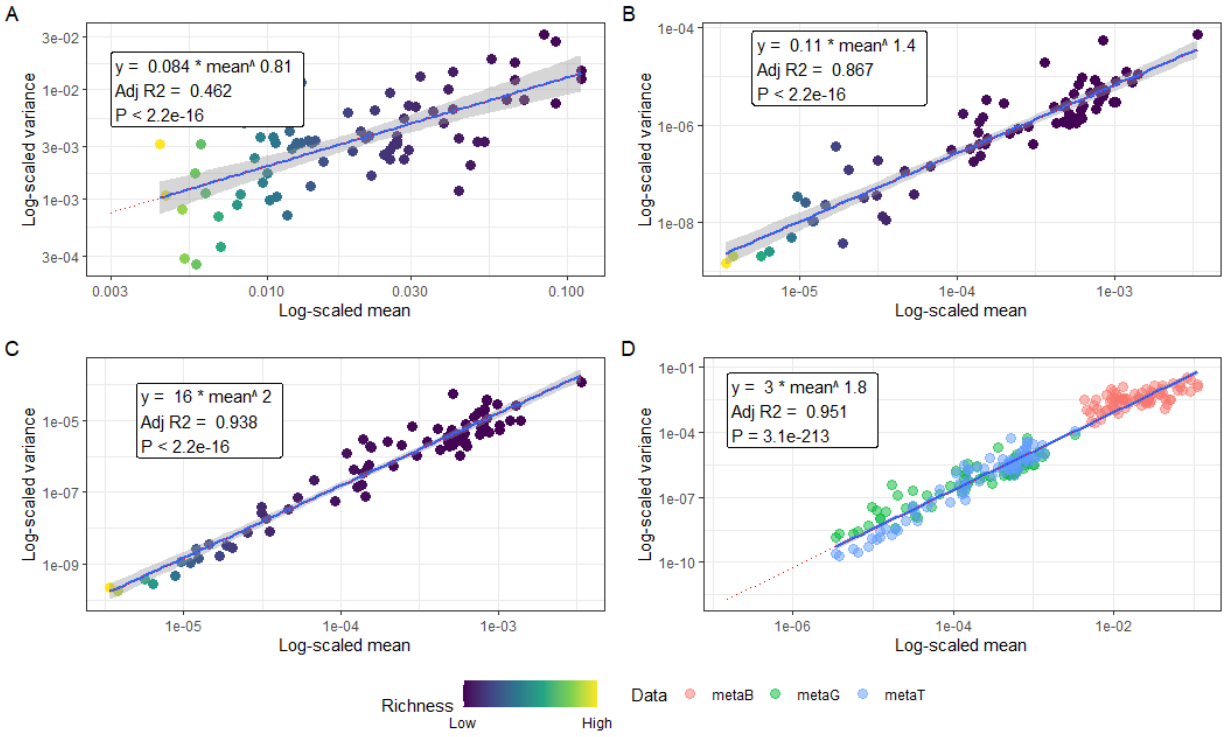

Figure 2.6. Scaling law relationship between the first and the second moment of the distribution of diatom communities detected using A) metabarcoding OTUs and B) metagenomic and C) metatranscriptomic unigenes. A simultaneous visualization of the three dataset is shown in D). The mean and variance are plotted on a log-log scale. Coefficients and exponents of scaling equations are shown for each linear regression.

Results of the analysis that related the third and fourth moment of the distributions (Fig. 2.7) showed the quadratic relation between skewness and kurtosis and confirmed the presence of a right tail for all three distributions. The most asymmetric stations, i.e., the ones with highest skewness, were also the ones with

longest tails. It also emerged that communities of OTUs were in general more symmetrical and less tailed compared to communities described by unigenes.

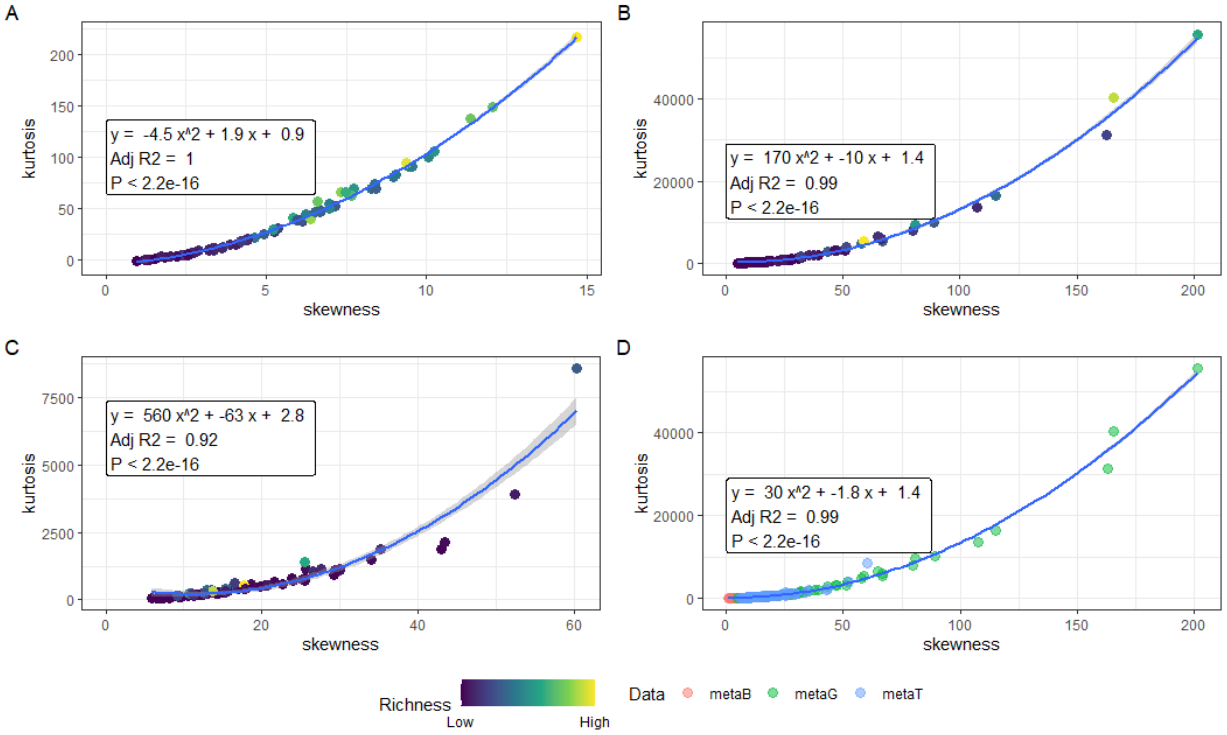

Figure 2.7. Quadratic relationship between the third and the fourth moment of the distribution of diatom communities detected using A) metabarcoding OTUs and B) metagenomic and C) metatranscriptomic unigenes. A simultaneous visualization of the three dataset is shown in D). Coefficients and of equations are shown for each polynomial regression.

2.3.5 Scaling of biodiversity indices with total abundance

Higher total number of reads (N) in a sampling station led to an increase in OTUs richness, dominance and rarity and a decrease in OTUs evenness. Rarity, evenness, and dominance scaled across four orders of magnitude in N (Fig. 2.8.). The scaling was allometric for all the indices considered, except dominance, that scaled nearly isometrically with N (z = 1.1). All the indices analysed showed significant high correlations with N.

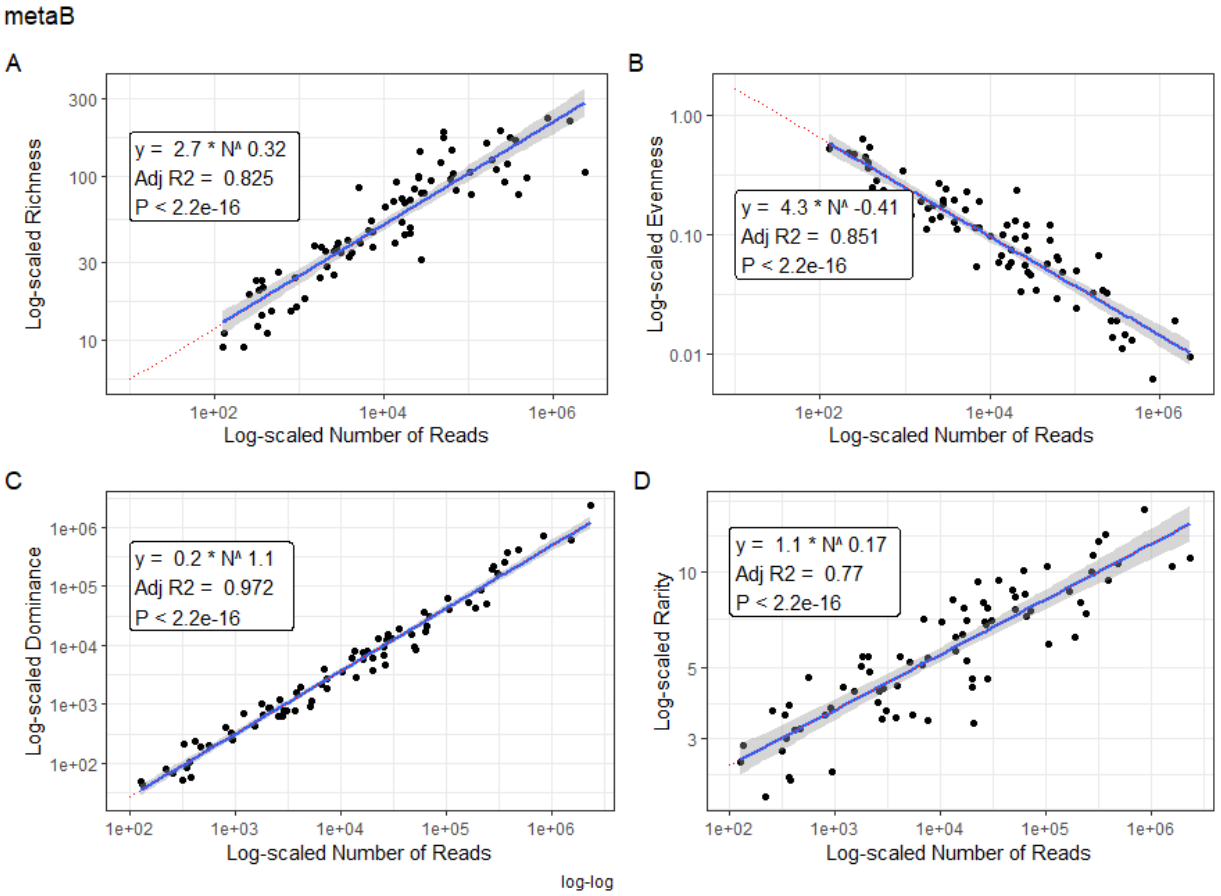

Figure 2.8. Scaling law relationship between number of OTU reads of each diatom community and the biodiversity indices of A) richness, B) evenness, C) dominance and D) rarity. Values are plotted on a log-log scale. Coefficients and exponents of scaling equations are shown for each linear regression.

Metagenomic unigenes showed a pattern similar to the one of OTUs and overall agreed with what was recently observed for both micro- and macro-organisms (*Locey and Lennon, 2016*): higher values of total abundance in one sampling station resulted in higher unigene richness, dominance, and rarity and a decrease in evenness (Fig. 2.9). However, besides richness, the analysed indices showed weaker correlations with N compared to metabarcoding.

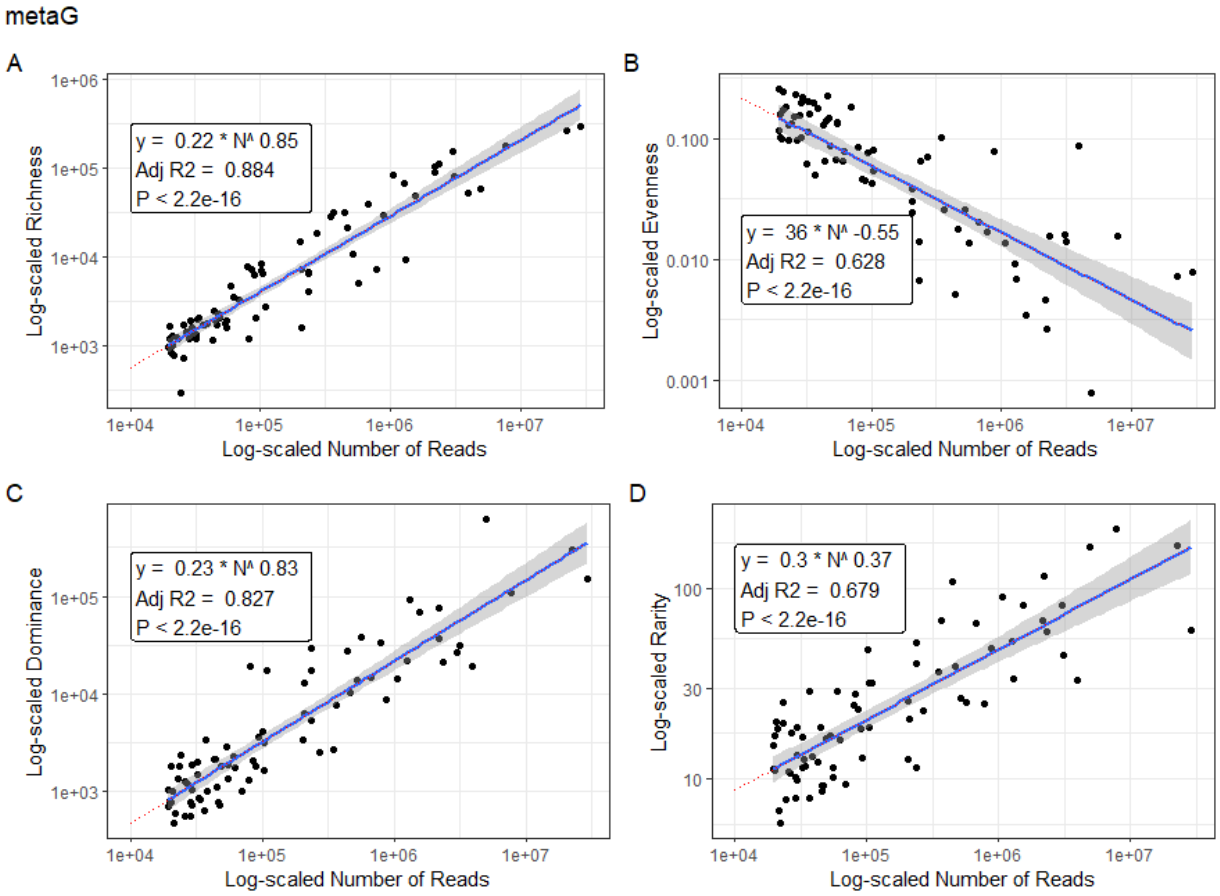

Figure 2.9. Scaling law relationship between number of metagenomic unigene reads of each diatom community and the biodiversity indices of A) richness, B) evenness, C) dominance and D) rarity. Values are plotted on a log-log scale. Coefficients and exponents of scaling equations are shown for each linear regression.

The results provided by unigenes in metatranscriptomes show trends that often deviated from the ones detected using DNA markers (Fig. 2.10). The most strongly maintained relationship was the one between N and the only biodiversity index that does not take into account the abundance information, i.e., richness. Richness, which by construction is identical for metagenomics and metatranscriptomics, scaled with N with a rate ($z = 0.7$) intermediate to the metabarcoding and metagenomic ones ($0.32 < z < 0.85$). Dominance scaled with N with a trend similar to metabarcoding and metagenomic data, but the points showed a higher dispersion from the regression line. Rarity showed a very weak positive relationship with N, while evenness increased with N, showing an opposite trend compared to

the one detected for metabarcoding and metagenomics. However, the correlation with evenness and rarity was weak (R^2 = -0.0118 and R^2 = 0.0157).

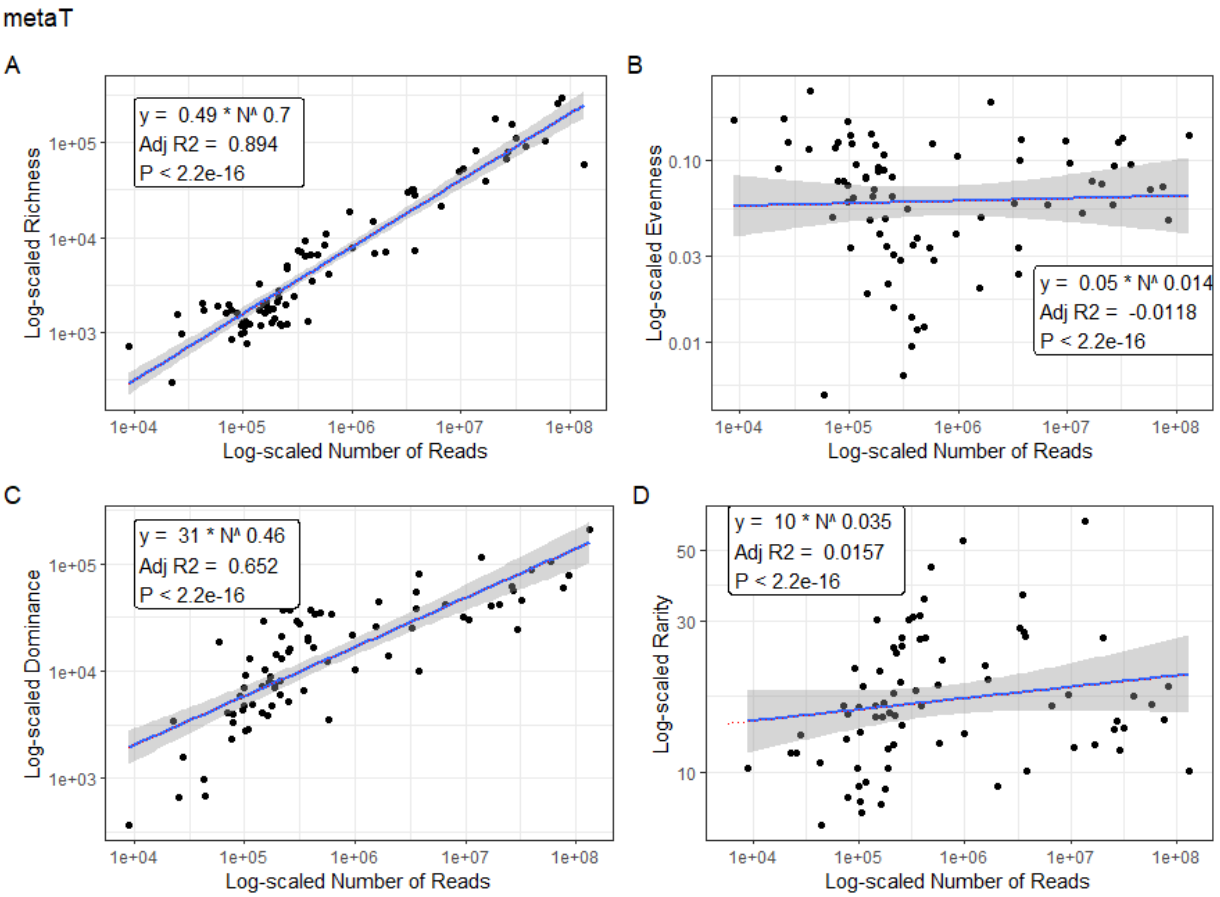

Figure 2.10. Scaling law relationship between number of metatranscriptomic unigene reads of each diatom community and the biodiversity indices of A) richness, B) evenness, C) dominance and D) rarity. Values are plotted on a log-log scale. Coefficients and exponents of scaling equations are shown for each linear regression.

2.4 Conclusion

The present chapter aimed at revealing emergent patterns in diatom macroecology by looking at abundance and distribution of biological units obtained through DNA-based (metabarcoding and metagenomics) and RNA-based (metatranscriptomics) techniques.

The geographical distribution of abundance and richness of diatoms showed similar trends within the three used datasets: diatoms are abundant at high

latitudes, where they often represent the most abundant taxon in the whole plankton community, and in cold waters in general. This pattern reflects in the high negative correlation between richness values and the physical parameters of temperature and salinity, as polar regions are the coldest and least salty waters of the globe, and this is especially true in the Arctic Ocean. When looking at metagenomic and metatranscriptomic richness, diatoms appear to hold a higher biodiversity also when nitrate and phosphate, crucial macronutrients for their metabolism, are highly concentrated. In *Tara* dataset, highest concentrations of these two nutrients are found especially in the Southern Ocean.

When comparing single communities, the three used data sets showed an overall similar structure made of few abundant and many rare biological units (both OTUs and unigenes). This pattern confirms what has been already shown for plankton inhabiting the global ocean (*Ser Giacomi et al., 2018*), where seascape features and oceanic currents are thought to enhance the presence of rare species (*Villa Martín et al., 2020*). Moreover, the Taylor's law, which describes the variance of a population as a power-law function of the mean population density, is verified for both OTUs and unigenes. Although it has been shown that Taylor's law can arise as an artifact in several conditions (e.g., *Cohen and Xu, 2015*), the fact that the sample average abundance spans across several orders of magnitude suggests that the observed relationship is a true result (*Grilli, 2020*). Moreover, the main results of this chapter were overall confirmed when using both a more relaxed and a more stringently filtered data set (see section 2.2.2), suggesting that the found relationships were robust to the inclusion or exclusion of the rarest entities.

In general, metabarcoding and metagenomic data showed very similar patterns. This is not unexpected, as both are based on the sequencing of DNA fragments; however, while metabarcoding technique targets a specific region of organisms' genomes, in this case the V9 region of the 18S rDNA gene, metagenomic information is obtained by processing a higher amount of information deriving by the random fragmentation of the genome content. This means that in the latter

case genes under different selective pressures are included. Metatranscriptomic data often departs from the patterns observed for both DNA-based markers. In particular, the difference between high-richness and low-richness sampling stations when looking at the rank abundance distributions (RADs) of genes suggests the occurrence of different mechanisms underlying their community assembly. Whether the observed pattern is the result of neutral and niche processes shaping the assembly of expressed genes or simply derives from technical biases is still to be understood; an ongoing effort to further characterize this behaviour is currently being performed by Emanuele Pigani (Stazione Zoologica Anton Dohrn, Italy), with the application of statistical fits to RAD curves and the investigation of the nature of the derived parameters that better describe the observed patterns, also taking into account the spatial location and the oceanographic condition of each sample. In addition to what was seen for RADs, the scaling law observed when looking at the relationship between the total abundance of a station (N) with its richness, evenness, dominance and rarity, was not always preserved for expressed genes. The more consistent pattern was the power law between the total abundance and total richness of a station, a relationship that does not surprise since the filtering approach used retained the same set of genes for metaG and metaT in a station; similarly, the abundance of the most expressed gene (i.e., metatranscriptomic dominance) scaled with N in the same way as it did for both metabarcoding and metagenomic data. On the other hand, the relationship between N and evenness and rarity was not verified. Contrary to richness and dominance, evenness and rarity are strongly affected by the distribution of the abundance of rare units. The observed relation with N did not suggest a specific behaviour, but it was rather defined by weak values of correlations and an undefined structure. The simultaneous decrease of evenness and increase of rarity, observed for both metabarcoding and metagenomic data when N increased, indicates that a greater abundance of units in a station reflects in the increment of rare entities that overall amplify the disparity between the dominant and non-dominant components of the community structure. On the

other hand, the abundance of a gene in the metatranscriptomic data not only reflects the number of species holding that gene in the community, but is also consequence of the regulation of its expression, a finely-tuned process that responds to environmental abiotic and biotic factors which every single individual is subject to. It is hard to derive general laws on functional units from the obtained result, but it is probable that the transcription regulation has important role in the observed patterns. However, since general trends can be observed for functional data too, it remains to be understood whether some components of the expressed genes are governed by general laws and how the contribution of the transcription shapes the properties of community structure of functional units. A further issue to consider is whether we can discriminate between a real bio-ecological signal and a non-homogeneous sampling effort. Macroecology often starts from the assumption that everything is equal everywhere and that, therefore, diversity increases with sampling effort. In metabarcoding and meta-omics studies, differences in sampling effort occur at several steps, from sample collection to sequencing, and depend on many reasons, including the volume of water collected and the way the DNA or RNA is distributed in the sampled space (*Grey et al., 2018*). Several approaches have been developed to adjust information on read counts during data processing, especially when comparing data from different projects; one example is rarefying the abundance values and then use the rarefied values as weights for random sampling, a procedure that often brings the counter effect of yielding too limited number of represented species that compromises the statistical robustness of analysis. Furthermore, the importance of niche-based processes cannot be neglected in this scenario, and it is clear that the environmental context, including temperature, nutrients, as well as biotic interactions, plays a role in determining the ultimate structure of species communities, and that sampling effort is not the unique explanation for the observed patterns.

The emergence of universal laws in ecology, observed for micro- as well as for macro-organisms, through the analysis of data from different projects and thus differently processed, is fascinating; the characterization of these laws will be crucial in understanding the interplay of neutral and ecological forces in shaping the observed patterns of abundance and diversity of communities across space and time. In conclusion, the results of this chapter strongly suggests that the integration of functional data from metatranscriptomic study offers the unique chance to explore these fundamental laws adding the perspective of genes that are actively expressed in the community.

Chapter III

Diatom community composition and iron metabolism from the North Atlantic to the Arctic Ocean

3.1 Introduction

The Arctic Ocean is the smallest of the world's five oceans; measuring about 14 million square kilometres, it represents around the 4% of the global ocean area (*Jakobsson, 2002*). Almost landlocked and surrounded by shallower channels, it is also referred to as the Arctic Mediterranean (*Jakobsson and Macnab, 2006*). It is composed by two main basins, the Eurasian and the Amerasian, separated by the Lomonosov Ridge, a submarine band of continental crust that extends across 1700 km and with a mean depth around 2000 m (*Timmermans et al., 2005*); the two basins are surrounded by the continental shelf area that makes up more than 50% of the total Arctic Ocean area, making it the shallowest of all the oceans, with an average depth of 1200 m (*Jakobsson et al., 2002*). In addition to being the northernmost, the smallest and shallowest, the Arctic Ocean is characterized by three main features: the high river discharge, the presence of sea ice, and strong seasonality.

In fact, although representing only the 1% of the world's ocean volume, the Arctic Ocean receives the 10% of global river discharge (*Aagaard and Carmack, 1989*), with six rivers acting as the main freshwater and organic matter sources (*Holmes et al., 2012*; *McClelland et al., 2012*; both reviewed in *Timmermans and Marshall, 2020*). The high rate of freshwater input makes the Arctic Ocean highly stratified, with a predominant salinity-driven stratification, contrary to the temperature-stratified subtropical oceans; as a result, a strong halocline separates the surface, low-

density and fresher waters from the dense, saltier deep waters (*Timmermans and Marshall, 2020*, and references therein).

Another peculiarity of the Arctic Ocean is the presence of sea ice, that influences the whole ecosystem through the regulation of light availability, heat exchange with the atmosphere (*Perovich and Polashenski, 2012*), surface temperature and salinity patterns, that in turn control stratification, mixing and nutrient concentrations (*Korhonen et al., 2013*). Sea-ice dynamics follow the strong seasonality that governs the whole Arctic ecosystem: in long dark winters, sea ice extends across the whole Arctic Ocean with an average thickness of 2 m (*Timmermans and Marshall, 2020*), reaching its maximum in March. With rivers freezing and the consequent reduced discharge, this season is characterized by a lower stratification and a higher mixing with nutrient replenishment (*Kohronen et al., 2013*). Sea ice then melts and breaks up during spring and summer until it reaches its minimum in September (*Polyak et al., 2010*).

The main inflows to the Arctic Ocean are the waters from the North Pacific and the North Atlantic Oceans (Fig. 3.1.), whose exchanges with the polar ocean are overall driven and structured by the stratification patterns (*Rudels et al., 2013*). Pacific waters enter the Arctic Ocean through the Bering Strait, providing an important source of fresher and warmer water (*Woodgate et al., 2010*; *Haine et al., 2015*). However, the largest water mass exchange occurs with the Atlantic Ocean via the Fram Strait and the Barents Sea Opening (see *Timmermans and Marshall, 2020*, and references therein); here the warm and salty water coming from the Gulf Stream flows as North Atlantic drift that crosses the Greenland-Scotland Ridge and reaches the Nordic Seas (*Hansen et al., 2008*). From the Norwegian Sea, it branches off towards a western and an eastern current. The western branch enters the Arctic Ocean through the Fram Strait that, with its 2600m depth, represents the deepest connection between the Nordic Seas and the Arctic Ocean. Fram strait is also 450km wide and holds two contrasting hydrographic patterns, separated by a frontal system (*Paquette et al., 1985*): while the West Spitsbergen Current (WSC)

brings warm, salty and nutrient-rich Atlantic Water into the Arctic Ocean on the eastern side of Fram Strait, the East Greenland Current (EGC) flows south out of the Arctic Ocean along the western side of Fram Strait (*de Steur et al., 2014*), and constitutes the major outflow area of Arctic water (*Beszczynska-Möller et al., 2012*) and sea ice (*Kwok et al., 2004*).

Atlantic Water also enters the Arctic Ocean from the Nordic Seas via the Barents Sea Opening (*Ingvaldsen et al., 2002*; *Schauer et al., 2002*), with a heat transport probably higher than the one occurring through Fram Strait (*Timmermans and Marshall, 2020*, and references therein). These currents are not only responsible of carrying warmer and fresher or saltier waters into the Arctic Ocean, but they also bring nutrients. For instance, a recent study (*Whitt, 2019*) emphasized the key role of the Gulf Stream in transporting macronutrients along the North Atlantic Ocean, with implications for the poleward nutrient advection and thus for global-scale patterns of biogeochemical cycles. The North Atlantic inflow, whether it crosses the Fram Strait or the Barents Sea Opening, is impacted by the air-sea fluxes and sea ice melting that lead to an overall cooling and freshening of the uppermost waters, with the warmest waters located deeper (e.g., *Rudels et al.,1996*; *Untersteiner, 1988*; both reviewed in *Timmermans and Marshall, 2020*).

Figure 3.1. Arctic Ocean Currents Map. Warm waters (red arrows) enter the Arctic Ocean via the Fram Strait and the Barents Sea Opening from the Atlantic Ocean and through the Bering Strait from the Pacific Ocean. Cold waters (blue arrows) circulate through the Arctic Ocean and flow back into the North Atlantic through Baffin Current. Slightly modified from AMAP, Icelandic Marine Research Institute, http://library.arcticportal.org/id/eprint/1494

The Arctic region is a focal point of climate change: due to complex feedback processes, it is experiencing the most rapid environmental changes on Earth (*Meredith et al., 2019*), with predictions picturing a future seasonally ice-free Arctic, where a thinner and more mobile sea-ice pack is present in winter/spring and absent in part of summer/fall (*Timmermans and Marshall, 2020*). The loss of Arctic sea ice is of particular concern given its key role in the regulation of planet's climate (*Euskirchen et al., 2013*; *CAFF, 2015*); sea ice is in fact tightly linked to all the aspects related to Arctic changes, from the temperature rise to the water freshening and

the changes in water stratification profile and circulation dynamics (*Timmermans and Marshall, 2020*, and references therein).

The temperature rise is strongly related to the heat transport from the main input waters; for instance, the Pacific Ocean water inflow increased its temperature and flux volume by 60% during 2001–2014 (*Woodgate, 2018*). However, a major role in the progressive Arctic ice reduction is played by the Atlantic Water inflow, whose temperature at both Fram Strait and the Barents Sea Opening is around 1-1.5 °C higher than the decadal means (*Muilwijk et al., 2018*). The progressive expansion of the Atlantic domain, with the consequent sea ice reduction, stratification weakening, and enhanced Atlantic Water flux into the Arctic Eurasian Basin is referred to as Atlantification of the Arctic Ocean (*Årthun et al., 2012*; *Polyakov et al., 2017*; *Lind et al., 2018*; *Timmermans and Marshall, 2020*), a phenomenon that concerns not only the abiotic component, but also the biotic portion of the whole Arctic ecosystem. For instance, it promotes the intrusion of Atlantic species into the Arctic Ocean, like in the case of the temperate coccolitophore *Emiliania huxleyi* (*Winter et al., 2014*), whose poleward expansion, originally imputed to sea warming, was recently attributed also to advective transport (*Oziel et al., 2020*). Bioadvection mechanisms thus play a key role in species range shift; however, not all the temperate species entering the Arctic Ocean are going to survive (*Wassman et al., 2015*), and the degree of their ecological success in the harsh polar environment, depending on factors ranging from species adaptation potential and plasticity to biotic interactions, is hard to predict.

More generally, the significant sea-ice cover decrease in the Arctic Ocean in terms of concentration, volume and duration has increased light availability for phytoplankton (*Arrigo and van Dijken, 2011*; *Bélanger et al., 2013*); as a consequence, the annual spring bloom is occurring earlier (*Kahru et al., 2011*) and ending later (*Lafond et al., 2019*), with some regions experiencing a second bloom in autumn (*Ardyna et al., 2014*), and an overall increase in total primary production from phytoplankton has been observed (*Arrigo and van Dijken, 2015*). Nevertheless, light is

not the only factor regulating primary production at high latitudes, and nutrient availability is considered similarly important or even more decisive than light (*Krisch et al., 2020*; *Tremblay and Gagnon, 2009*; *Schourup-Kristensen et al., 2018*). In the Arctic Ocean, primary productivity is mainly driven by nitrate availability (*Tremblay and Gagnon, 2009*; *Codispoti et al., 2013*; *Taylor et al., 2013*; *Tremblay et al., 2015*; *Schourup-Kristensen et al., 2018*; *Randelhoff et al., 2020*); however, a role for iron limitation has also been shown in some regions (*Aguilar-Islas et al., 2007*; *Taylor et al., 2013*; *Lannuzel et al., 2016*). It is known that iron constitutes the limiting nutrient for phytoplankton growth in ~30–40% of the oceans (*Moore et al., 2001, 2004*; *Cohen et al., 2017*), including the sub-polar North Atlantic (*Nielsdóttir et al., 2009*; *Ryan-Keogh et al., 2013*; *Moore and Doney, 2007*; *Moore et al., 2009*; *Achterberg et al., 2018*), a region that hosts the largest annual phytoplankton bloom in the global ocean (*Behrenfeld et al., 2019*), therefore representing a major component of the oceanic carbon cycle (*Achterberg et al., 2018*). Moreover, a N-Fe co-limitation has been suggested in the eastern Fram Straight (*Nielsdóttir et al., 2009*; *Rijkenberg et al., 2014*; *Browning et al., 2019*; *Krisch et al., 2020*), crossed by the warm salty West Spitsbergen Current (WSC).

In this context, understanding how organisms react to nutrient limitation can provide insight into what could happen in the future, as highlighted by a recent experimental study on iron responses of the two cosmopolitan diatom genera *Pseudo-nitzschia* and *Thalassiosira* (*Cohen et al., 2017*). Here, authors showed how species that live in Fe-limited habitats adopt molecular and physiological strategies that, while variable across taxa and environments, are strictly linked to N and C metabolism. This conclusion is particularly relevant in the context of the predicted fluctuations of iron in the Arctic Ocean. Here, in fact, the availability of this micronutrient is again linked to sea ice, whose seasonal melting, with the following iron release, is considered a critical trigger of phytoplankton growth (*Lannuzel et al., 2016*, and references therein). The future scenario of iron availability in the Arctic Ocean is complex: the seasonal iron supply provided by sea ice will decrease in a warmer ocean, while its solubility will increase as a consequence to ocean

acidification (*Millero et al., 2009*), a process that, on the other side, is also expected to reduce the effectiveness of the redox processes involved in iron uptake by phytoplankton (*Shi et al., 2010*), thus ultimately reducing the overall Fe bioavailability (*Taylor et al., 2013*). In conclusion, it is still uncertain whether iron will limit phytoplankton growth in a warmer and acidified Arctic Ocean and the implications to the Arctic primary production.

Today, this primary production is mainly attributed to diatom-dominated spring blooms (*Sakshaug, 2004*; *Perrette et al., 2011*), that also support a cascade of processes at higher trophic levels and play a major role in carbon sequestration and export (*Tremblay et al., 2002*; *Wassmann et al., 2008*). Diatoms, mainly represented by the genera *Chaetoceros* and *Thalassiosira*, are the dominant blooming group in many regions of the Arctic Ocean (*Balzano et al., 2017*, and references therein), together with the prymnesiophyte *Phaeocystis pouchetii* (*Lafond et al., 2019*, and references therein) and coccolithophores (*Ardyna and Arrigo, 2020*), and are particularly important in coastal locations (*Booth and Horner, 1997*; *Lovejoy et al., 2002*).

However, in a warmer Arctic with a longer sea-ice-free season, mixotrophic eukaryotes, like dinoflagellates, are predicted to be favoured over more strictly phototrophic phytoplankton like diatoms, especially during the long dark periods (*Ardyna and Arrigo, 2020*), a replacement that could impact both the phenology and the magnitude of Arctic blooms (*van de Poll et al., 2019*), also with the risk of promoting the proliferation of harmful bloom-forming species (*Ardyna and Arrigo, 2020*, and references therein). Diatom contribution to primary production is predicted to decline in the future at global scale, including the Arctic Ocean, with the only exception of the Southern Ocean, either with links to species range shift (*Tréguer et al., 2018*) or to local adaptation processes (*Krause and Lomas, 2020*). Nevertheless, future trends in nutrient concentration and oceanographic patterns are likely to significantly vary across the distinct Arctic regions (*Lafond et al., 2019*), and different results have been shown in predictive studies on climate change responses from phytoplankton in various regions of the Arctic. For instance, while

diatom blooms in central Baffin Bay are not expected to be particularly affected by climate change (*Lafond et al., 2019*), this might not be the same in the Barents Sea, whose transition from a cold stratified environment to a warmer and mixed "Atlantified" regime is supposed to occur in the near future (*Lind et al., 2018*). Whether or not diatoms will still dominate the spring productivity in a warmer Arctic Ocean is yet to be fully understood (*Krause and Lomas, 2020*), but it is clear that the implications of this phenomenon will potentially affect the whole marine ecological dynamics, including productivity rate, biotic interactions, population mixing and harmful species occurrence (*Ardyna and Arrigo, 2020*, and references therein).

In order to predict the consequences of the environmental changes in the Arctic phytoplankton community, it is thus necessary to study its taxonomic and functional biodiversity and reveal who are the actual components of these assemblages, where they occur, and whether they are or not equipped with the molecular repertoire for coping with the expected changes. This baseline will also allow us to identify future changes (*Egge et al., 2021*), with important implications; for instance, a recent study based on metabarcoding data collected in the Beaufort Sea (*Balzano et al., 2017*) revealed how the majority of diatoms had a specific northern/polar distribution, like in the case of *Pseudo-nitzschia arctica* and the *Chaetoceros neogracilis* species complex. This high level of endemism in the Arctic Ocean (*Terrado et al., 2013*) is a key factor to be considered in conservation efforts, since climate change is likely to favour cosmopolitan generalist species at the expense of endemic taxa that are strictly adapted to this extreme habitat. Finally, most studies are focused on data from the Arctic region. Given the continuous inflow of Atlantic water masses (transporting also species into the Arctic), a proper assessment of the Arctic ecosystem dynamics and resilience instead requires a characterization of the role of the North Atlantic current system populations, if any.

Our knowledge of species distribution and diversity in the Arctic Ocean is still scarce (*Lovejoy, 2014*; *Egge et al., 2021*), but recent meta-omic studies suggested it could be an emergent biodiversity hot spot, e.g., for viruses (*Gregory et al., 2019*). In

this context, the aim of this chapter is to investigate the taxonomic and functional landscape of diatom communities along the oceanic currents that flow from the Gulf Stream to the Arctic Ocean crossing the North Atlantic Ocean and reaching the Pacific-influenced Arctic Chucki Sea. In particular, I will assess the taxonomic composition and similarity between diatom communities populating the surface of *Tara* Oceans and *Tara* Oceans Polar Circle sampling stations using metabarcoding data and analyse it in the context of the emerging environmental gradients and oceanographic currents. Metatranscriptomic data will then be integrated for a functional exploration focused on diatom iron metabolism: in order to exploit the unprecedented meta-omics data produced by the *Tara* Oceans expeditions I here selected specific genes involved in iron uptake, transport and storage, and used a robust statistical and phylogenetic pipeline to map the distribution and abundance of target genes across the currents of the studied transect.

3.2 Materials and Methods

3.2.1 Study Site

A total of 24 *Tara* Oceans sampling stations were selected, with 11 stations belonging to the North Atlantic Ocean and 13 to the Arctic Ocean, sampled during three seasons (Fig. 3.2). In agreement with the main objective of the study, all the stations of the North Atlantic Ocean were included, except one (station 142) located in the central Gulf of Mexico; station 143 is placed at the beginning of the Gulf Stream while the majority of North Atlantic stations are located in the North Atlantic Gyre, from station 144 to 152, with the only exception of station 145 that is mainly composed by the cold waters flowing from the Labrador Sea. Stations 155 and 158, located along the North Atlantic Drift, are thus right at the gateway from North Atlantic to Arctic Ocean; station 163 is almost at Fram Strait, while station 168 is in the Barents Sea, the region currently more influenced by the

Atlantification phenomenon. We also included six samples from the Kara-Laptev Sea (stations 173, 175, 178, 180, 188 and 189), with station 189 separated by a land mass from the rest of the stations in the same area. We excluded from the analysis stations belonging to the Labrador Sea and Canadian Archipelago, since they mainly represent outflows from the Arctic Ocean, but considered one (station 191) in the East Siberian Sea and three stations of the Chukchi Sea (from 193 to 196), all influenced by Pacific Ocean inflow. For the selected stations, surface samples, i.e., samples collected in the first 10 m of the water column, were analysed.

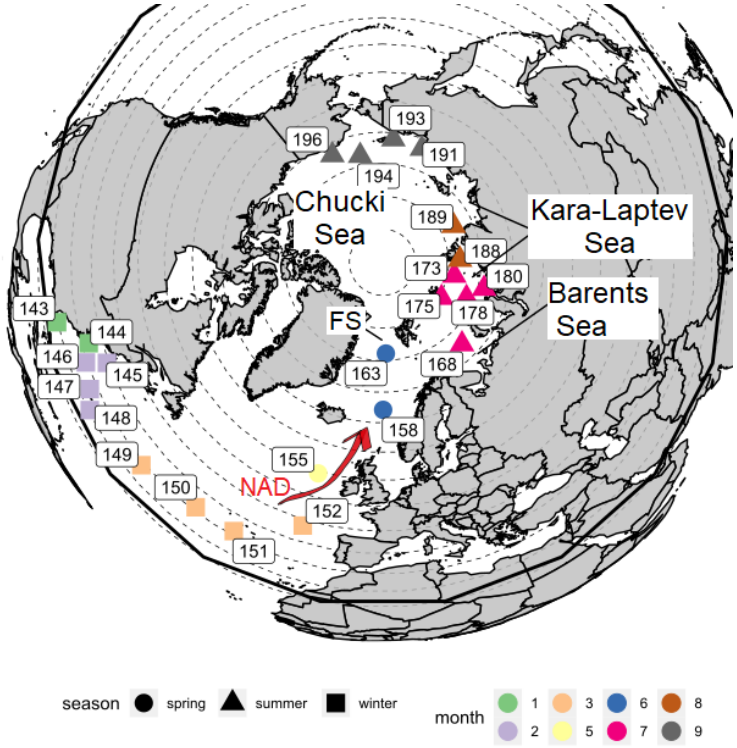

Figure 3.2. Localization and seasonal information of stations included in the present study. NAD: North Atlantic Drift; FS: Fram Strait.

3.2.2 Environmental characterization

Environmental data considered for this study come from three sources: *in situ Tara* Oceans data, values extracted from the World Ocean Atlas 2018 (WOA18; *Boyer et al., 2018*) and from the up-to-date PISCES model (*Aumont et al., 2015*). I used three

physical variables: Temperature, Salinity and Sea Sunshine Duration, together with the information on month and season of sampling. The selected macronutrients were nitrate, phosphate and silicate; micronutrients such as cobalt, copper, iron, manganese and zinc were also included. Values of the same parameter deriving from the three different sources were compared and correlated to each other: to compensate for missing physicochemical values in the *Tara* Oceans *in situ* data set, climatological data of WOA18 were ultimately preferred. The final list of environmental descriptors selected is shown in Table 3.1.

Table 3.1. Environmental descriptors selected for this study. Values extracted from WOA18 and PISCES are monthly mean surface (5 m depth) values.

Parameter	Symbol	Unit	Source
Temperature	Temp	°C	WOA18
Salinity	Sal	PSU	WOA18
Sea Sunshine Duration	SSD	Min	Astronomical Applications of the U.S. Naval Observatory
Nitrate	NO_3^-	µmol/L	WOA18
Phosphate	PO_4^{3-}	µmol/L	WOA18
Silicate	$Si(OH)_4$	µmol/L	WOA18
cobalt	dCo	picomol/L	PISCES2
copper	dCu	nmol/L	PISCES2
iron	dFe	nmol/L	PISCES2
manganese	dMn	nmol/L	PISCES2
zinc	dZn	nmol/L	PISCES2

Scaled values of the selected environmental descriptors (obtained by *standardizing* each of the variables to mean equals to 0 and standard deviation to 1) were used to build a Principal Component Analysis (PCA) on sampling stations. The PCA was performed through the R package *vegan* (*Oksanen et al., 2013*). A geographical representation of the main groups identified by the PCA was constructed and

plotted, together with box plots showing the distribution of values of the environmental factors in the clusters defined by the PCA.

3.2.3 Community structure

Operational Taxonomic Units (OTUs) obtained after swarm clustering (*Mahé et al., 2014*) of metabarcoding reads representing the V4 and V9 regions of the 18S rDNA gene were used to study the main patterns in diatom alpha and beta diversity and the taxonomic composition of stations. All the three main size fractions of diatoms have been considered, i.e., the nano-, micro- and meso-plankton, respectively representing the size ranges 3-20 µm (or 5-20 µm), 20-180 µm and 180-2000 µm.

3.2.3.1 Alpha diversity

Richness, expressed as number of different OTUs in a sample, and abundance, expressed as total number of reads in a sample, have been calculated in each station. These measurements have been performed both using V4 and V9 information and related to each other. A filtering on V9 data was applied in order to reconcile the information provided by the two metabarcoding markers; in particular, OTUs occurring with less than 10 reads in each sample and occurring in less than two stations were discarded. Pearson's correlations between the values of richness and abundance obtained through the two different metabarcoding markers was computed both before and after the applied filter.

3.2.3.2 Cluster analysis

Diatom communities belonging to different stations were compared and grouped using a hierarchical clustering approach separately on the V4 and the filtered V9 dataset. In particular, I used a Ward D2 clustering of the Jaccard similarity matrix calculated on presence-absence data of OTUs. Subsequently, I calculated the absolute numbers of OTUs exclusive to the two main emerging clusters and the amount of OTUs shared by the two. Once these three pools of OTUs (exclusive to

one cluster, shared, exclusive to the other cluster) were identified, I measured the relative contribution of each pool to the overall composition of each station, providing a general characterization of the main inter-cluster an intra-cluster patterns.

3.2.3.3 Simplex approach

With the aim to analyse the reasons behind diatom community structure from the North Atlantic to the Arctic Ocean, I then adopted the approach proposed by *Podani and Schmera (2011)*, that presented an intuitive and powerful conceptual and methodological framework for comparing communities through the simultaneous calculation of different biodiversity indices and the evaluation of the relative contribution of the measured ecological phenomena to the observed patterns. This approach has been widely utilized in community ecology and also recently extended to metacommunity studies (*Podani et al., 2018*). These authors proposed to represent data structure as point cloud in a triangular plot built upon three main biodiversity indices: The Jaccard similarity (S) and the decomposition of its complement, the Beta diversity (Jaccard dissimilarity), into its two main components, i.e., richness difference (D) and species replacement (R). Given two sites *j* and *k*, with *a* representing the number of shared species between *j* and *k*, *b* expressing the number species present only in *j* and not in *k*, and *c* indicating the species present only in *k* and not in *j*, the three indices can be defined as follows:

- The Jaccard similarity coefficient (S) is one of the most widely used indices in community ecology, and represents the size of the intersection between two sites (i.e., the shared species) divided by the union of the sites (i.e., all the species from the two sites):

$$S_{jk} = \frac{a}{(a+b+c)} \qquad (1)$$

Its complement can be decomposed into two additive indices (*Podani and Schmera, 2011*; *Carvalho et al., 2012*): richness difference and species replacement;

- Richness difference (D) is the component of beta diversity only ascribed to changes in richness between the two sites j and k:

$$D_{jk} = \frac{|\, b - c \,|}{(a + b + c)} \qquad (2)$$

Its complement, richness agreement, represents the largest fraction of species for which the two sites are equally rich in species;

- Species replacement (R) is the component of diversity linked to the substitution of species among sites:

$$R_{jk} = \frac{2\, min\{b, c\}}{(a + b + c)} \qquad (3)$$

Its complement is the nestedness, that measures the condition of species of a site being a subset of species in another site.

The three indices Sjk, Djk and Rjk sum to 1; therefore, they can be used to represent the data structure as an equilateral triangle, the SDR-simplex diagram (Fig. 3.3.), where the relative position of each point represents the pairwise comparison between two sites *j* and *k* respect to the three vertices S, D and R, each of one corresponding to situation when one value is 1.0 and the other two are 0.0. In this ternary plot the proximity of a point to a given vertex is proportional to the respective coefficient value; in this way, it is possible to graphically discern the relative importance of each of the selected aspect of biodiversity in shaping the

observed patterns. The analysis was performed with base R functions, while the ternary plots were constructed with the *ggtern* package (*Hamilton, 2016*).

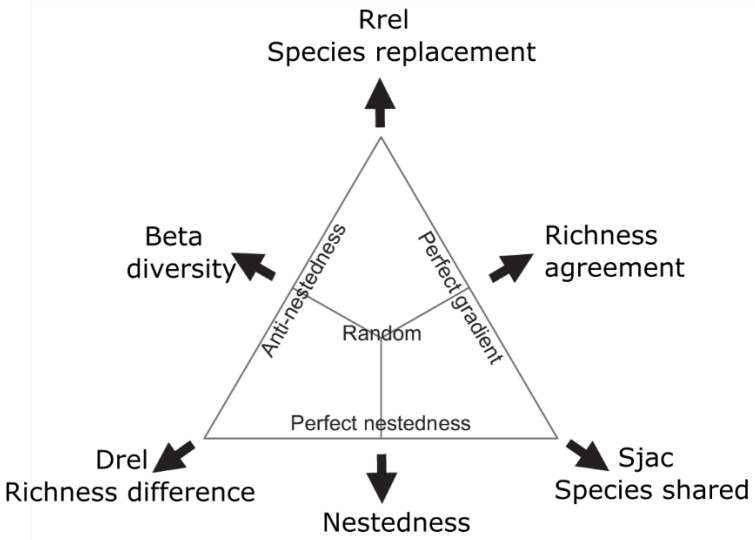

Figure 3.3. Representation of ecological concepts in the SDR-simplex diagram. Slightly modified from Podani et al. (2011).

3.2.3.4 Taxonomic composition

The abundance of each diatom genus was calculated relatively to the total diatom abundance for each sampling station, depth, and size class.

Genus-level composition assessments are widely performed in ecological studies on plankton, although the major variation between individuals would be more precisely addressed if looking at species level. Annotation at the species level was incomplete for both V4 and V9 *Tara* Oceans metabarcoding data; as diatoms are a highly specious group, the lack of sufficient taxonomic resolution in community ecology studies might be problematic. However, the aim of the study was to describe the overall diatom community composition in order to obtain an estimation of the contribution of the main diatom groups to the communities inhabiting the North Atlantic and Arctic Ocean stations sampled. Therefore, the

genus-level composition investigation was sufficient for the scope of the present study.

All the genera whose relative abundance in a sample represented at least the 10% of the total diatom abundance were individually plotted; otherwise, their abundance was summed and they were collectively indicated as "other" genera. The constructed bar plots showed the taxonomic composition at the genus level of each sample. This calculation has been done separately for V4 and V9 markers, and the two obtained plots were compared.

3.2.4 Iron metabolism

The pipeline presented here, included in the following sections (3.2.4.1 - 3.2.4.2), has been originally developed by Dr Camilla Borgonuovo at Stazione Zoologica Antohn Dohrn in the context of the study described in Chapter VI of the present thesis. I integrated this methodological approach and applied it also here, with a few modifications. A schematic view of the used pipeline is shown in Fig. 3.4.

Figure 3.4. Schematic representation of the pipeline used to select and filter the target genes.

3.2.4.1 Sequences search

Representative and well annotated diatom protein sequences corresponding to genes involved in iron metabolism were selected from literature, in particular from two recent review papers that summarised the role of these genes in phytoplankton (*Behnke and LaRoche, 2020*) and diatoms (*Gao et al., 2021*). I selected a total of 10 iron-related genes, involved in the uptake, transport and storage of this trace element. A complete description of the genes and their functions is provided in Table 3.2.

Table 3.2. Iron metabolism genes selected for the study.

Protein	Process	Subprocess	Function
NRAMP (Natural Resistance Associated Macrophage Protein)	Iron uptake and transport	general divalent metal uptake and transport	Fe^{2+} uptake; Fe^{2+} intracellular transport
ZIP (IRT-ZIP: Iron Regulated Transporter-Zinc Transporter)			
ISIP2 (Iron Starvation Induced Protein 2)		Phytotransferrin-mediated Fe^{3+} uptake and transport	phytotransferrin; Fe^{3+} uptake
ISIP1 (Iron Starvation Induced Protein 1)		siderophore-mediated Fe^{3+} uptake	endocytosis-mediated siderophore transport
FBP (Ferrichrome Binding Protein)			siderophore binding protein
FRE (Ferric Reductase)		Reductive high affinity Fe^{3+} uptake	Fe^{3+} reduction
MCO (MultiCopper Oxidase) - FET (Ferroxidase)			Fe^{2+} oxidation
FTR1 (FeIII permease)			Fe^{3+} permease
FTN (FerriTiN)	Iron storage	Storage/ homeostasis	iron storage/iron homeostasis
ISIP3 (Iron Starvation Induced Protein 3)			putative storage

Multiple protein sequence alignments for each gene were carried out using MUSCLE algorithm (*Edgar, 2004*) implemented in the software MEGAX, version 11.0.8 (*Kumar et al., 2018*), and then transformed into Profiles Hidden Markov Models (HMM) using the command hmmbuild as part of the HMMer package, version 3.1b2 (*Finn et al., 2011*). The HMM profiles were used as queries for sequence similarity searches using the HMMER algorithm against the MATOU-v2 dataset (Marine Atlas of *Tara* Oceans unigenes; *Carradec et al., 2018*), i.e., the gene catalogue used for the macroecology-based analysis described in Chapter II. Here this database was consulted through the Ocean Gene Atlas website (OGA; http://*Tara*-oceans.mio.osupytheas.fr/ocean-gene-atlas/; *Villar et al., 2018*) by leaving the default threshold on the *e*-value, i.e., the number of expected hits of similar quality that could be found just by chance; in particular, only sequences with an e-value lower than $1E^{-10}$ were selected.

Notwithstanding the use of annotated sequences to build the HMM profiles and the imposed threshold on the *e*-value, a considerable number of false positives is expected to be obtained by searching the gene catalogue. Therefore, the output hits were furtherly filtered according to their taxonomic assignation and the associated bit scores. In particular, retrieved unigenes were grouped according to their phylum annotation into four categories: Bacillariophyta, Other Stramenopiles, Other Taxa, Unknown. Genes belonging to Other Taxa were excluded from the analysis, while for the other groups two different bit score thresholds were chosen. In particular, genes belonging to the phylum Bacillariophyta were retained only if associated with a bit score ≥ 50, as this value has been shown to be statistically significant for inferring homology in protein alignments (*Pearson, 2013*). A more stringent criterion was imposed on genes belonging to Other Stramenopiles or of unknown annotation, where the bit score threshold was set as 100. At this point I removed all the unigenes that, although passing the so far imposed filters, were never present at the selected sampling stations, depth and size classes. The obtained set of sequences was considered for the subsequent analysis.

3.2.4.2 Phylogenetic analysis

The applied filtering procedure based on taxonomy and bit score annotation allowed a first screening of the results. In order to get insight into the phylogenetic relationships between sequences and thus further select the most probable diatom sequences involved in the functions of interest, I applied a phylogenetic approach. First, for each gene, the selected MATOU-v2 sequences were aligned with the diatom queries used to build the original HMM profiles; in order to achieve a better resolution of the relationship between queries and results, homologous sequences of each target gene held by non-diatom species were included. They were collected either from *Behnke et al. (2020)* and/or downloaded from the database PLAZA Diatoms 1.0 (https://bioinformatics.psb.ugent.be/plaza/versions/plaza_diatoms_01/).

In particular, for each gene, I made a Blastp search in PLAZA with the sequences used to build the HMM profiles as queries. I then selected the output genes belonging to the same orthologous gene family of the used queries, as each annotated gene family contains sets of genes that are all orthologous one another, i.e. genes diverging through speciation from a common ancestral gene (*Gabaldón and Koonin, 2013*) and thus likely holding the same biological function. The integration of already characterized orthologous genes from non-diatom species was aimed at helping the assessment of functional and taxonomical annotation from of the unknown MATOU-v2 unigenes.

Multiple sequence alignments were then performed using the MUSCLE algorithm (*Edgar, 2004*) implemented in the software MEGAX, version 11.0.8 (*Kumar et al., 2018*) followed by automatic removal of poorly aligned positions with trimAl v1.4. (*Capella-Gutierrez et al., 2009*) (gap threshold -gt 0.1). Maximal likelihood phylogenetic trees with 1000 ultrafast bootstrap replicates were constructed using IQ-TREE v2.1.3 (*Nguyen et al., 2015*), with automatic ModelFinder substitution model. Trees were visualized and edited with FigTree version 1.4.3 (*Rambaut and Drummond, 2012*); the

editing was relative to the layout of trees and gene IDs of sequences that were manually renamed with their taxonomic assignation. Moreover, poorly supported branches containing MATOU-v2 sequences that did not cluster with any of the known diatom sequences were collapsed. According to tree topology and bootstrap support values, I finally visually and manually selected the final set of MATOU-v2 unigenes that more robustly represented the target iron-metabolism genes in diatoms.

3.2.4.3 Occurrence of markers across stations

After obtaining the catalogue of genes that underwent bit score, taxonomy and phylogenetic-driven selection, I retrieved the information on the presence of each MATOU-v2 gene in each sample of interest. I then built a heatmap showing the presence and absence of each transcript in each station; both genes and stations were grouped through Ward D2 clustering based on the Jaccard similarity indices calculated on the occurrence matrix.

3.2.4.4 Distribution of transcripts across diatom taxa

The taxonomic assignation of each transcript was explored; in particular, I constructed bar plots showing the taxonomic annotation of genes at the class level, in order to distinguish between centric and pennate diatoms, also evaluating the fraction of genes of unknown annotation. When the genus annotation was present, I separately explored the percentage of transcripts annotated as centric or pennate genera for each iron gene.

3.2.4.5 Gene expression from the North Atlantic to the Arctic Ocean

Unigene abundance is expressed as “percent of total genes per sample”, that is gene read coverage in RPKM (Reads Per Kilobase covered per Million of mapped reads) divided by the sum of the total gene coverage for the sample. So, the abundance results are the fractions of homologs of all genes in the whole sample.

To compare the relative expression of the different iron genes, I scaled the abundance of transcripts to the total abundance of iron metabolism genes per each sample (i.e., station and size class) and taxonomical group. This normalization allowed the comparison of expression of genes among different samples, but it is obviously relative to the whole diatom iron genes present in the sample. Bar plots showing the relative abundance of transcripts belonging to the different markers were displayed for each station, also accounting for differences in size classes and taxonomic assignation. In particular, I selected the 3-20/5-20 and the 20-180 µm size classes and separately plotted the information of relative expression of genes for centric and pennate diatoms. Moreover, I grouped genes according to the process in which they are involved, in order to provide an overview of the relative expression of the main iron metabolism processes across stations from the North Atlantic to the Arctic Ocean. Gene expression was visualized together with the concentration of iron and nitrate in each station. Ideally, the comparison of the abundance of genes in metagenomic data and their expression in metatranscriptomic data would have allowed a more in-depth investigation of the observed patterns. However, when selecting for each sample only unigenes in common between the two meta-omic data sets, a not negligible number of genes was lost from the expression set. This limit did not allow the comparison between the two data.

3.3 Results and Discussion

3.3.1 Environmental characterization

Comparisons and Pearson's correlations between values of the same parameter deriving from the *Tara in situ* data, climatological data of WOA18 and from PISCES model are shown in Fig. 3.5 and Table 3.3. The correlations for the two physical parameters, i.e., temperature and salinity, were very solid among the three data sets. When looking at macronutrients, some differences arose, like the case of high

PISCES-derived nitrate values in the Arctic Ocean compared to the other two sources of data. These discrepancies can relate to errors in the model that can be amplified at low concentrations. The reason of the observed pattern could rely to the fact that for a number of years PISCES model produced a high level of nitrate in the Arctic due to the river outflow (Dr. Alessandro Tagliabue, personal communication). On the other hand, there was not a complete *Tara in situ* data set for nitrate, but only for the total nitrogen (i.e., the sum of nitrate and nitrite). Moreover, *Tara in situ* data did not cover all the target stations for phosphate and silicate. Given the high correlation values for temperature and salinity, and in order to compensate for missing physicochemical values in the *Tara* Oceans *in situ* data set, climatological data of WOA18 were ultimately selected for the subsequent analysis.

Figure 3.5. Line plot showing the comparison among TARA Oceans in situ data (blue line), PISCES data (green line) and WOA18 (red line) values for the physical parameters and the main macronutrients selected along the study site. Units of measure for each parameter are the same as in Table 3.1. "Nitrogen" stays for "NO_3^-" in PISCES and WOA18 and for "NO2NO3" in Tara.

Table 3.3. Values of pairwise Parson's correlation, p-value and 95% confidence interval between Tara in situ data, WOA18 data and PISCES data, as calculated for the physical parameters and the main macronutrients selected along the study site.

Env. factor	Comparison	Pearson cor.	p-value	95% C.I.
Temperature	TARA-WOA	0.99	< 2E-16	0.975; 1.00
	PISCES-WOA	0.99	< 2E-16	0.98; 1.00
Salinity	TARA-WOA	0.8	5E-06	0.58; 0.91
	PISCES-WOA	0.97	5E-16	0.94; 0.99
Nitrate	TARA-WOA	0.62	0.003	0.25; 0.83
	PISCES-WOA	0.06	0.780	-0.35; 0.45
Phosphate	TARA-WOA	0.76	7E-05	0.48; 0.89
	PISCES-WOA	0.75	3E-05	0.49; 0.88
Silicate	TARA-WOA	0.43	4E-02	0.03; 0.72
	PISCES-WOA	0.91	2E-10	0.82; 0.97

The PCA constructed upon the physical and nutrient factors considered explained 76% of total variability among stations (Fig. 3.6A). The first axis was the most representative (52% of explained variation) and was mainly related to salinity, iron and manganese (Fig. 3.6C). The variable that mostly contributed to the second axis (24% of explained variation) was zinc (30% of contribution), followed by copper (22%) and sunshine duration (Fig. 3.6C).

However, looking at the PCA biplot it seems it produced an artifact known as the Horseshoe Effect, in which the second axis is curved and twisted relative to the first, not representing a true secondary gradient. The interpretation of the horseshoe in a PCA biplot is that data have a single dominant gradient. When performing the scatterplot of salinity versus zinc, i.e., the main contributors to PC1 and PC2 (not shown), the points were arranged in a way similar to what occurred with the PCA in Fig. 3.6. This means that the rotation of vectors of parameters that

best represented the data was the one that projected the points on the salinity - zinc gradient.

The PCA biplot allowed a first visualization of groups and outliers. Looking at the results we could identify three main groups, whose distribution of environmental parameters is shown in box plots in Fig. 3.7.

The first group was composed only by North Atlantic Ocean (NAO) stations characterized by high temperature and high salinity (Fig. 3.6A and 3.7). The only NAO stations that did not group within this cluster were stations 145, 152 and 155. These three were in fact part of the second cluster, that included stations that belong to the "Atlantified" part of the Arctic Ocean (see section 3.1) and the transition between the two basins, together with the station 145 that hosts the cold waters flowing from the Labrador Sea (see section 3.2.1). This cluster was overall characterized by concentrations of NO_3^- , Cu and Zn much higher than the other two groups (Fig. 3.6A and 3.7). A third cluster was composed by inner Arctic Ocean

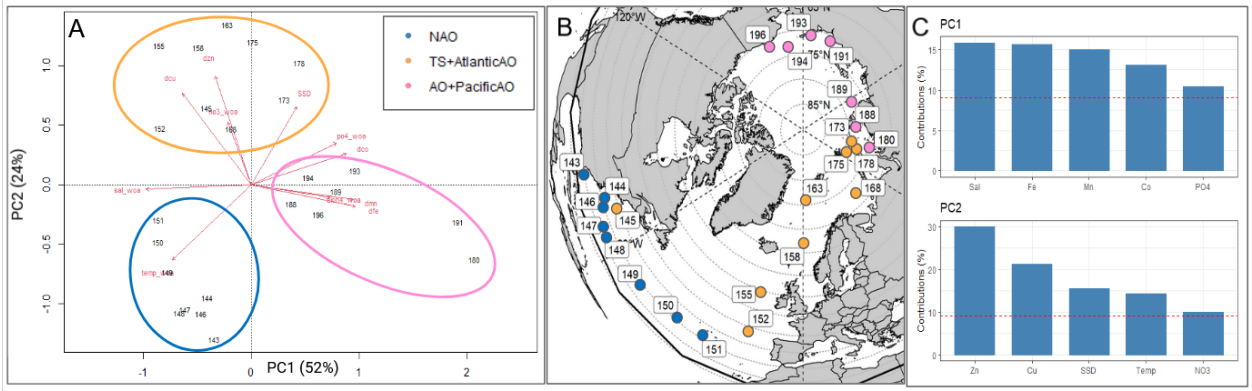

Figure 3.6. A) Principal component analysis of environmental parameters associated with each station. Circles represents the three clusters identified by eye (NAO: North Atlantic Ocean; TS + AtlanticAO: Transition + Atlantic-Arctic Ocean; AO+PacificAO: Arctic Ocean + Pacific-Arctic Ocean). B) Geographical localization of stations of each PCA-identified cluster. C) Contribution of the top 5 variables to the first two principal components PC1 and PC2. The red dashed line represents the expected value if the contribution where uniform.

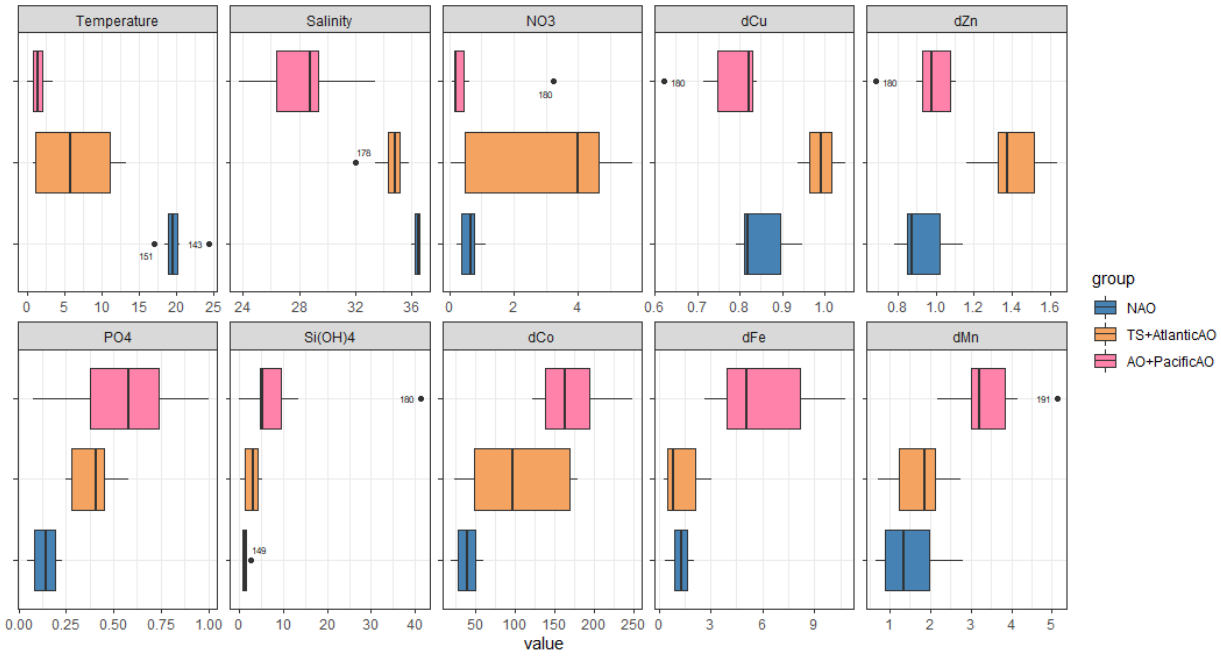

Figure 3.7. Distribution of environmental factors according to the clusters defined in the PCA. NAO: North Atlantic Ocean; TS: Transition; AO: Arctic Ocean.

(AO) stations as well as the ones that are more influenced by the Pacific Ocean inflow, i.e., stations 191-196. This group was characterized by high concentrations of PO_4^{3-}, Co, Fe and Mn. Two stations of this group, namely station 180 and 191 emerged as outliers from the PCA biplot, and the box plot allowed to detect which parameters they were outlier for: station 180 had the highest nitrate and silicate and the lowest Cu and Zn concentrations among the stations of its cluster, while 191 was characterized by extraordinary high Mn concentration.

3.3.2 Community structure

Patterns of richness and abundance across the study site have been analysed using the two most common metabarcoding markers for eukaryotes, i.e., the regions V4 and V9 of the 18S DNA (Fig. 3.8.). There are some stations for which V4 information was absent; in particular, station 178 for all size fractions, stations 143, 144, 146 and 152 for nanoplanktonic samples, while stations 163 and 191 were not represented for microplanktonic and mesoplanktonic diatoms. V9 showed

values higher than V4 both in the number of distinct Operational Taxonomic Units (OTUs), i.e., richness, and in the number of sequencing reads, which we used as a proxy for abundance of OTUs; in particular, V9 richness was around 6 times higher than V4, and its abundance exceeded the V4 values of more than one order of magnitude.

Figure 3.8. Bar plots showing the abundance, expressed as absolute number of reads, and richness, expressed as number of different OTUs, for each sampling station in the three different size fractions, as measured with the two metabarcoding markers V4 (blue) and V9 (red).

This higher amount of OTUs in V9 compared to V4, also observed in other studies (e.g. *Piredda et al., 2017*; *Tragin et al., 2018*) is counter-intuitive if we consider the differences in terms of length of the two regions: the V4, with its ~450bp length, is longer than the V9, that measures only 150bp. Reads obtained from the two markers are processed and clustered into OTUs with the same approach, i.e. using the swarm algorithm (*Mahé et al., 2014*); therefore, one would expect that V4 yield to a higher number of OTUs compared to V9. One explanation of the unexpected result relies to the observation that, despite being longer, nucleotide changes of V4 region are actually all concentrated in a small part of the region (*Monier et al.,*

2016, *Tragin et al., 2018*), and this could have led to a different level of nucleotide polymorphisms across the whole 18S region.

In order to better investigate this pattern, I applied a filter to V9 dataset retaining only OTUs that held at least 10 reads in each sample and occurring in at least 2 stations. Results on the Pearson's correlation coefficients between V4 and V9 richness patterns are shown in Fig. 3.9. When V9 was not filtered, correlations with V4 were high and significant for both microplanktonic and mesoplanktonic size classes (corresponding to the size ranges 20-180 and 180-2000 µm, respectively), while this relationship did not hold for the smallest diatoms (Fig. 3.9A). When applying the abundance and occurrence filtering thresholds, Pearson's correlations with the smallest size fraction became significant and high and the correlations between the two markers for the other two size classes improved. Moreover, the absolute numbers of richness values (values on x and y axis in Fig. 3.9) became more similar between the two datasets. This result suggests that rare OTUs detected through V9 metabarcoding are the ones causing the discrepancies in richness between the two 18S markers. Removing the fraction of rare OTUs in V9 increased the correlation on richness between the two markers and made their quantities comparable. The reason behind the higher detection of rare entities in V9 compared to V4 remains to be identified.

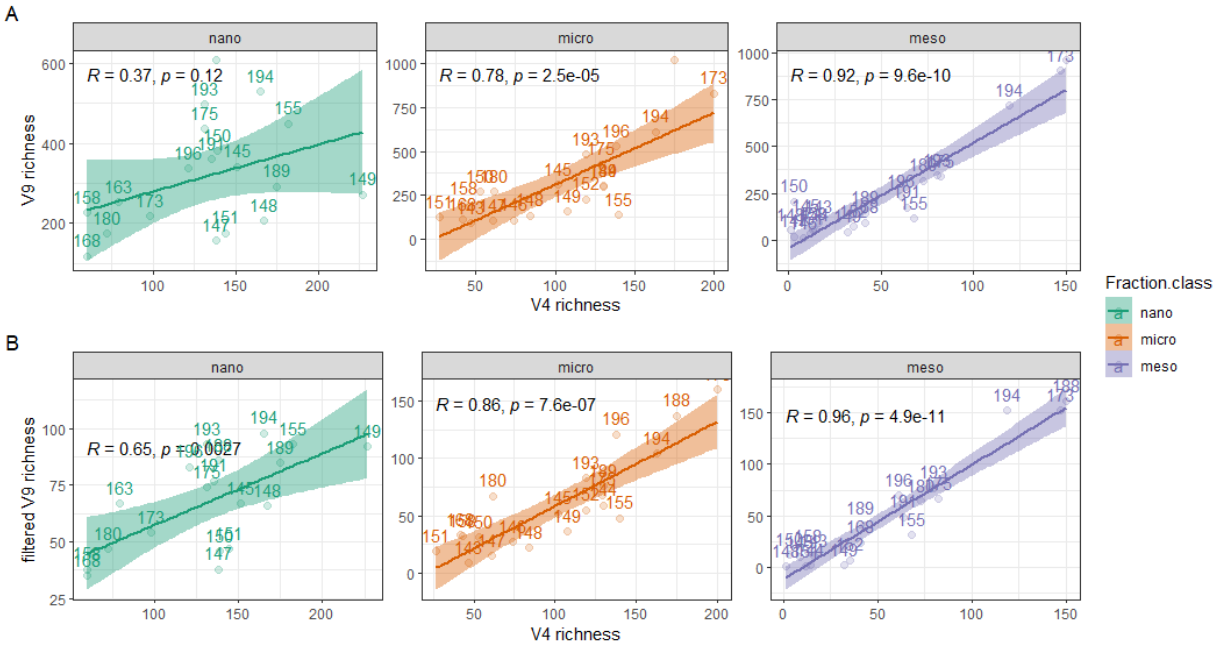

Figure 3.9. Relationship between diatom richness for the three size classes using V4 data and A) V9 data, B) V9 filtered data. Pearson's r coefficients and p-values are shown for each linear regression.

However, the same reconciliation between the values obtained with the two 18S markers did not occur when looking at the abundance values, thus suggesting that the relative abundance of all rare taxa together was still too low to make a difference (Fig. 3.10).

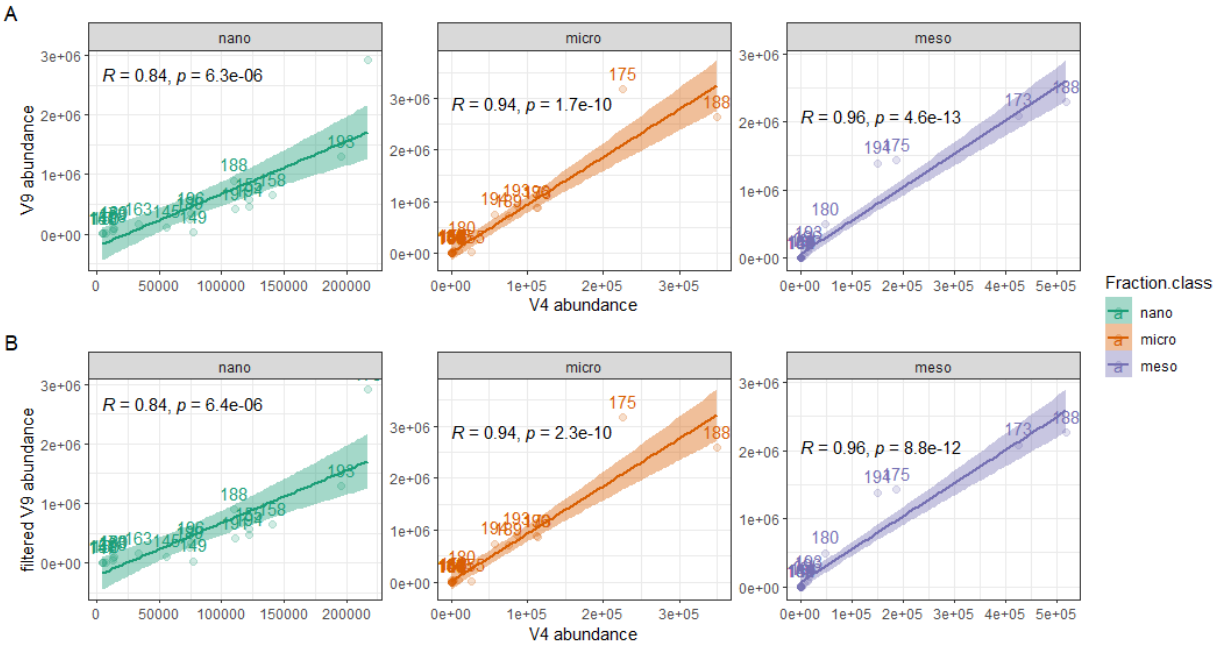

Figure 3.10. Relationship between diatom abundance for the three size classes using V4 data and A) V9 data, B) V9 filtered data. Pearson's r coefficients and p-values are shown for each linear regression.

Although different in terms of number of reads and OTUs, the two markers showed similar geographical trends, with an overall increase of diatom richness and abundance from the North Atlantic (NAO) to the Arctic Ocean (AO) (Fig.3.8).

This result apparently contrasts one of the most widely recognised macroecological patterns in nature, i.e., the Latitudinal Diversity Gradient (LDG, *Hillebrand, 2004*), that refers to the increase of species biodiversity from the poles to the tropics. However, it has been shown how this spatial configuration is not unequivocally respected by all organisms; diatoms, for instance, have displayed an inverse LDG in streams (*Soninen et al., 2016*), and a more complex pattern at global ocean scale (*Busseni et al., 2020*), being less subject to a general temperature-dependent latitudinal decrease of biodiversity. This is due to the role of abiotic or biotic factors other than latitude as main drivers of biodiversity (*Soininen and Teittinen, 2019*) and to the methods used to describe it. For instance, while *Ibarbalz et al. (2019)* found diatom communities showing LDG, *Busseni et al. (2020)* did not confirm this pattern using the same dataset; here one difference relies on the fact

that while the first authors measured diversity through Shannon index, a metric that describes how abundance of units is distributed, the second calculated richness as number of distinct units, in the same way I do in the present chapter, revealing how the two measures cannot be simply compared as they weight differently low abundance taxa, that make up a substantial and metabolically active portion of protist communities in the oceans (*Logares et al., 2014*), included polar regions.

Moreover, it has to be noticed that higher values of richness and abundance in the Arctic Ocean compared to the North Atlantic might rely on the fact that diatom Arctic communities have been sampled in coastal sites and in summer, while the North Atlantic stations are often more found in the open ocean, and sampled in winter, thus expecting less productivity. However, we did not find a higher richness in North Atlantic coastal sites compared to the open ocean samples, especially when looking at the North Atlantic- Arctic transition stations, sampled in spring in a region that hosts the largest annual phytoplankton bloom in the global ocean (*Behrenfeld et al., 2019*).

The observed increase in abundance of diatom OTUs in the AO was consistent across all size fractions, while richness values were more variable across different sizes. NAO stations were rich in nanoplanktonic diatoms as well as the AO communities; microplanktonic communities showed a pattern that similar to the small size fraction according to V4 and similar to the bigger size class according to V9. Large diatom richness strictly reflected what is shown for abundance (as observed in Chapter II): i.e., high values in the AO and much lower values in the NAO. The observed differences between NAO and AO, already shown in the bar plots in Fig. 3.8, were confirmed by the Ward D2 clustering of the Jaccard similarity matrix calculated on presence-absence data (Fig. 3.11). From this moment, unless specified, I kept the applied threshold used on V9 in order to obtain comparable results on the two datasets. The cluster analysis showed that both metabarcoding markers grouped stations in two main clusters, a North Atlantic and an Arctic one.

Although the internal distribution of stations varies between V4 and V9, the two main geographical clusters were composed by the same stations for both markers. According to V4 clustering, stations were distributed into small clusters that overall reflected the geographical distribution of stations, with the exception of station 145 that, although located next to the east US coast, clustered into a group of stations that represents the transition zone between NAO and AO; however, as described in section 3.2.1, the localization of this station makes it a particular case, as it is mainly influenced by cold water flowing from the northern Labrador Sea. The geographical grouping was less clear in the V9 clustering, but a NAO-AO transition clade was evident here too.

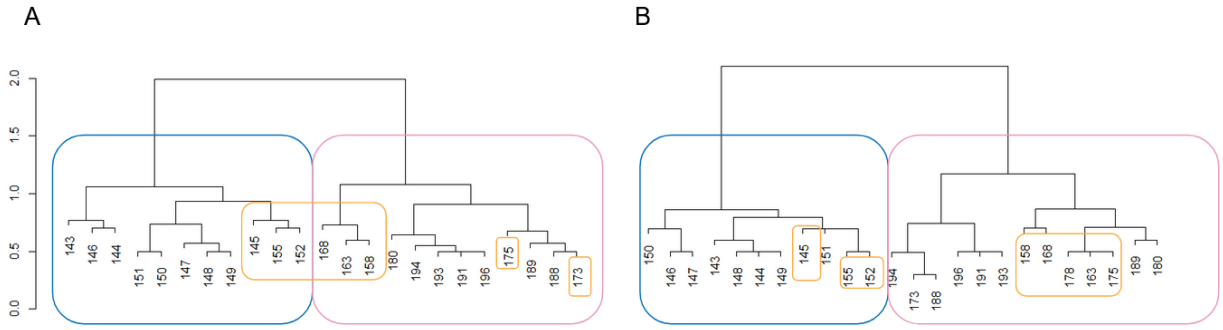

Figure 3.11. A. Comparison of diatom communities at the OTU level revealed by Ward D2 hierarchical clustering on Jaccard similarity matrices calculated on presence-absence data using A) V4 and B) V9 filtered data sets. Colours of rectangles group stations according to the environmental PCA groups (blue: NAO; orange: transition + Atlantified AO; pink: inner AO + Pacific-influenced AO; see Fig.3.6.)

Venn diagrams showing the number of OTUs that were unique to one or shared between the two main groups obtained through cluster analysis are shown in Fig. 3.12. The two clusters shared 21% of the detected OTUs and showed similarly high levels of unique genes according to V4 (NAO: 42%; AO: 37%), while V9 detected a higher amount of OTUs unique to AO compared to NAO (53% and 18% respectively), with 29% of OTUs shared. Although non unique, this result highlights the potentially endemic nature of a large part of diatom Arctic OTUs, suggesting geographic isolation or the presence of cryptic diversity.

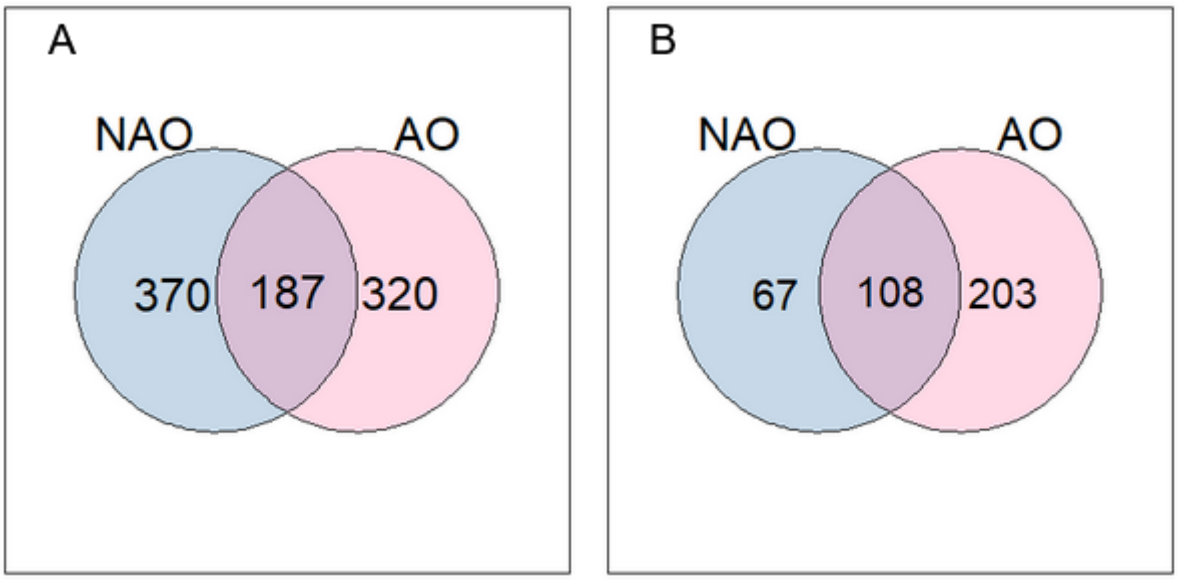

Figure 3.12. Venn diagrams of number of OTUs in NAO and AO basins according to A) V4 and B) filtered V9.

I then explored the contribution of each of the three pools of genes, i.e., unique to NAO, unique to AO and shared, to the total richness of a station (Fig. 3.13). The analysis of both markers showed a clear NAO-AO transition group composed by stations 155, 158, 163 and 168, that displayed the highest level of shared OTUs.

Figure 3.13. Bar plots showing the contribution of the three pools of genes to the total richness in each station according to A) V4 and B) filtered V9.

The simplex approach (*Podani and Schmera, 2011;* section 3.2.3.3) allowed displaying each pair of stations as a point in the triangular space whose precise positioning was obtained by the values of the three main indices calculated for each comparison, i.e., S (Jaccard Similarity), D (Richness Difference) and R (Species Replacement) (Fig. 3.14). The plot is coloured in agreement with the basins compared, and allowed a deeper understanding of the causes of the emerged difference between NAO and AO. Patterns of beta-diversity between the two basins were mainly dominated by species replacement, pointing to the fact that diatom communities at higher latitudes and extreme environmental conditions were not a mere subset of what flows from lower latitudes, and nestedness was never observed. Moreover, although a higher richness was found in the Arctic Ocean compared to the North Atlantic Ocean, richness difference (D) seemed not to be the main driver of the observed separation between the two basins. When looking at the intra-oceanic comparisons, both NAO and AO showed internal high variability. There was not a clear pattern emerging in NAO-NAO and AO-AO comparisons; this heterogeneity suggests that both basins showed an internal regionalization.

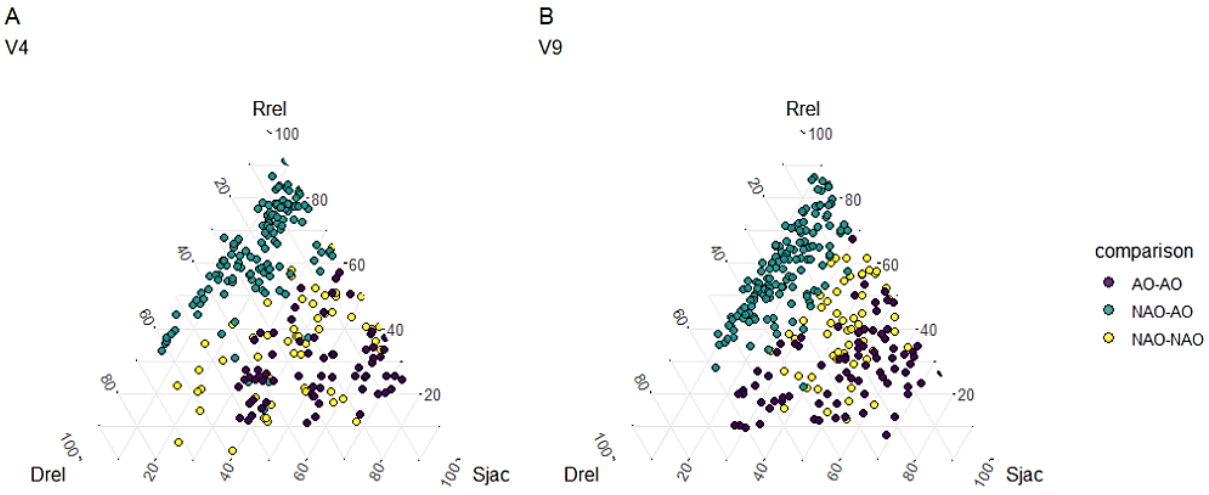

Figure 3.14. Ternary plots illustrating diatom community structure using the SDR-simplex approach (left: V4; right: V9). Letters at vertices refer to relativized measures (Sjac: relativized similarity; Drel: richness difference; Rrel: species replacement), otherwise to percentage contributions.

Taxonomic composition of diatom communities detected via V4 and V9 for each size fraction is shown in Fig. 3.15. The two metabarcoding markers showed similar patterns and provided a comparable picture of the distribution of diatoms analysed at the genus level, thus supporting the use of either V4 and V9 in global scale studies (*Tragin et al., 2018*).

V9 held an overall higher number of genera with unknown annotation compared to V4; V9 is shorter and this could lead to a higher uncertainty of taxonomic annotation. However, the correct annotation also strongly depends on what is present in available databases (*Tragin et al., 2018*). Both markers held a high quantity of genera annotated as "other"; these are genera, that, alone, contributed to less than 10% to the total abundance in a station. However, the cumulative abundance of these rare taxa can be predominant, and it is especially true in NAO, where also the relative amount of "unknown" was higher than in AO.

We could overall detect three diatom communities:

1. NAO community: it was overall dominated by rare genera and taxa of unknown annotation. The most abundant known diatoms in those stations were *Thalassiosira* spp., followed by *Chaetoceros* spp., *Coscinodiscus* spp., *Actinocyclus* spp. *Ceratulina* and *Guinardia* spp.;
2. Three transition stations (namely 155, 158 and 163): they were dominated by *Fragilariopsis* and *Pseudo-nitzschia* spp.;
3. AO community: it was overall dominated by *Chaetoceros* spp. Together with *Chaetoceros*, we noticed the local dominance of *Actinocyclus* and *Thalassiosira* and in the Atlantic-influenced Arctic and the high abundance of *Proboscia* spp. in the Pacific-influenced stations. Interestingly, *Proboscia* was never abundant in other stations, suggesting that it could actually be favoured in those stations after advection from the subpolar region of the North Pacific Ocean.

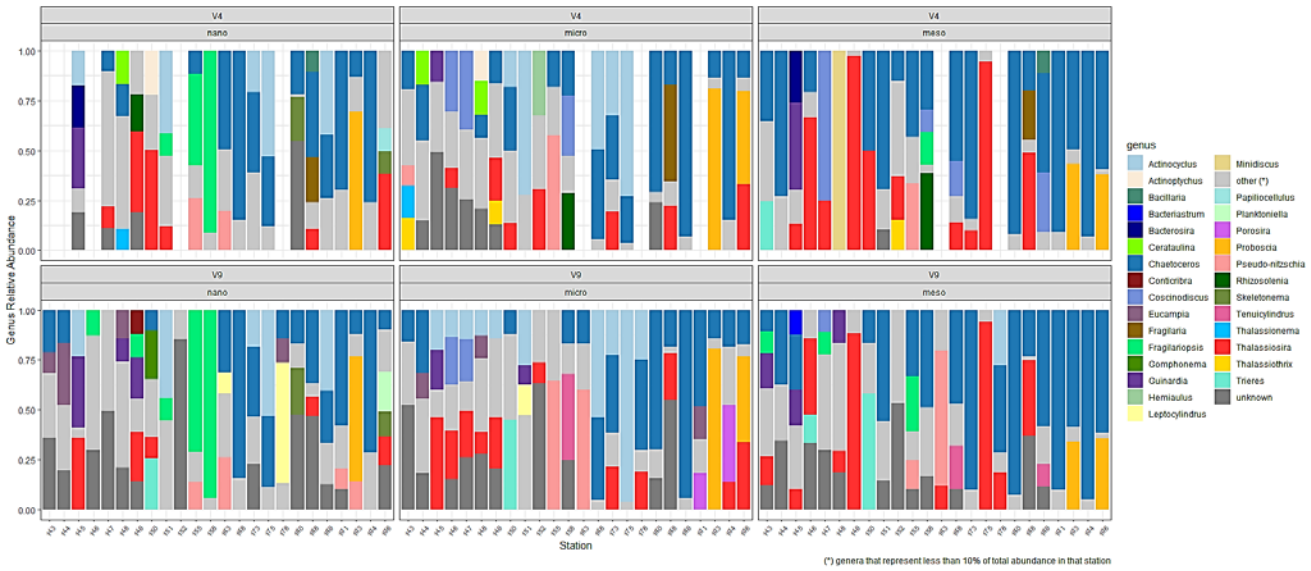

Figure 3.15. Bar plots representing the abundance of each diatom genus relatively to the total diatom abundance in V4 (top) and V9 (bottom) data, for each sampling station and size class. Nano: nanoplankton (3-20 or 5-20 µm); micro: microplankton (20-180 µm); meso: mesoplankton (180-2000 µm). Genera whose relative abundance in a sample represented less than the 10% of the total diatom abundance were labelled as "other".

Besides a general trend, some size-specific patterns emerged. One of the most interesting is what is found in the NAO-AO transition stations, in particular stations 155, 158 and 163: these were dominated either by *Fragilariopsis* spp. (that can reach up to ~90% of the total abundance) or by *Pseudo-nitzschia* spp., depending on the size fraction considered. In particular, *Fragilariopsis* appeared as the dominant genus for smallest diatoms, while *Pseudo-nitzschia* dominated the 20-180 µm size fraction. However, given the claimed difficulty of the 18S region as resolutive marker for phylogenetic studies of the two genera (*Lim et al., 2016*) and the morphological, ecological and genetic closeness between *Fragilariopsis* and *Pseudo-nitzschia* species (*Lundholm et al., 2002*), this result could either mean that the two genera were both highly favoured at the sampling conditions of the NAO-AO transition stations or that both 18S markers not always successfully differentiated between these two ecologically relevant raphid pennate genera. Finally, we observe how *Thalassiosira* spp. strongly dominated NAO stations in all size fractions, with a higher abundance in the large size fraction, with the striking case of station 149.

3.3.3 Iron metabolism

3.3.3.1 Sequences search

As discussed in section 1.2.4, iron uptake, transport and storage are complex processes that diatoms perform adopting many strategies with a variety of genes involved. Moreover, important differences between pennate and centric diatoms in the way they metabolize this trace element have been suggested (*Gao et al., 2021*, and references therein).

Representative and well annotated diatom protein sequences corresponding to genes involved in iron metabolism were selected from literature. In particular, 10 genes corresponding to 4 different strategies of iron uptake and transport and 2 genes involved in iron storage and general homeostasis were selected (see Table 3.2). As shown in Table 3.4., using a variable number of sequences ranging from a minimum of 5 to a maximum of 23, I retrieved a high number of hits from MATOU-v2 database (at global scale), from a minimum of 620 to a maximum of more than 18,000. No relationship between the number of sequences used to build the HMM profiles and the number of sequences retrieved from MATOU-v2 was found (Fig. 3.16).

Table 3.4. Iron metabolism genes selected for the study. The number of sequences used to build each HMM profile is indicated, as well as the total number of hits retrieved from MATOU-v2 through HMM search. Fcy: Fragilariopsis cylindrus; Pmu: Pseudo-nitzschia multistriata; Pgra: Pseudo-nitzschia granii; Tpse: Thalassiosira pseudonana; Toce: Thalassiosira oceanica; Ptri: Phaeodactylum tricornutum; Fso: Fistulifera solaris; Sro: Seminavis robusta; Pmse: Pseudo-nitzschia multiseries. Sya: Synedra acus; Ccri: Cyclotella criptica.

Gene	HMM profile sequences											hits retrieved
	Fcy	Pmu	Pgra	Tpse	Toce	Ptri	Fso	Sro	Pmse	Sya	Ccri	tot
NRAMP	3	1	0	1	1	0	0	0	0	0	0	6	3675
ZIP	1	2	0	0	2	1	2	0	0	0	0	8	4476
ISIP2	2	2	1	0	0	5	5	0	0	0	0	15	18212
ISIP1	2	4	0	0	2	1	2	0	0	0	0	11	1880
FBP	4	1	0	0	4	1	13	0	0	0	0	23	1391
FRE	4	1	0	1	2	2	5	0	0	0	0	15	4145
FET/ MCO	3	2	0	0	2	0	0	2	0	2	4	15	8316
FTR	1	1	0	5	0	0	4	0	0	0	0	11	620
ISIP3	2	0	0	1	2	1	3	0	0	0	0	9	5824
FTN	2	0	0	0	0	1	0	0	2	0	0	5	7728

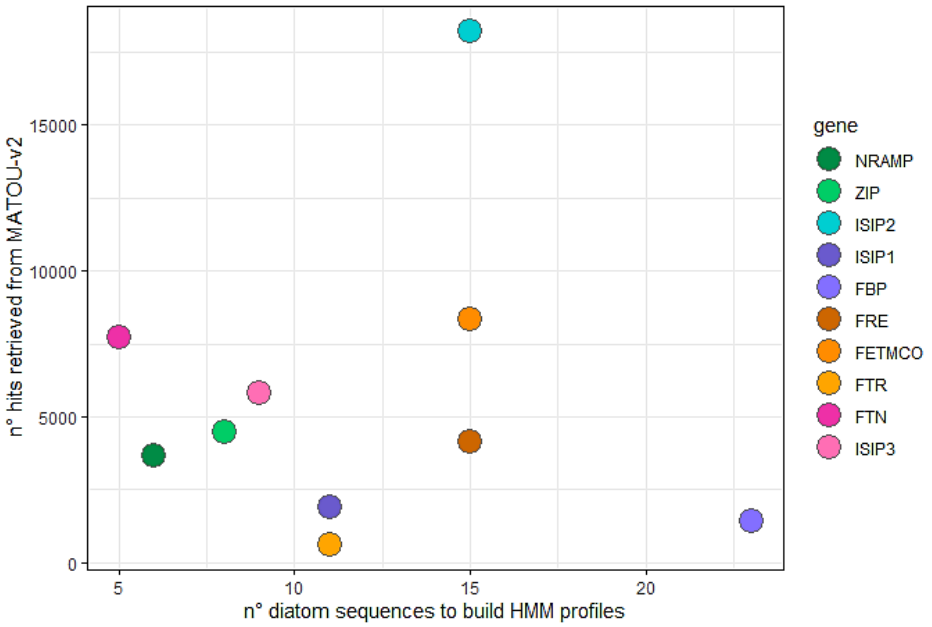

Figure 3.16. Scatterplot showing the relationship between the number of diatom sequences used as queries to build HMM profiles and the number of hits retrieved through HMM search from MATOU-v2 dataset.

The applied filtering pipeline based on bit score values of alignments and taxonomic annotation of sequences filtered out a variable number of genes found in MATOU-v2 dataset at global scale; the sequences were further selected retaining only genes found in the selected North Atlantic and Arctic Ocean stations. Numbers and percentages of retained genes are shown in Table 3.5.

Table 3.5. Progressive filtering according to bit score values, taxonomy and geographical distribution. Percentages are relative to the hits retrieved from MATOU-v2 through HMM search.

Gene	hits retrieved from MATOU-v2	1. bit-score + taxonomy filtering		2. geographical filtering	
		n° taken seq.	% taken seq.	n° taken seq.	% taken seq.
NRAMP	3675	847	23	532	63
ZIP	4476	695	16	529	76
ISIP2	18212	10303	57	6033	59
ISIP1	1880	1732	92	893	52
FBP	1391	1037	75	573	55
FRE	4145	1008	24	669	66
FET/MCO	8316	1501	18	997	66
FTR	620	443	71	356	80
ISIP3	5824	3217	55	1807	56
FTN	7728	293	4	189	65

3.3.3.2 Phylogenetic analysis

As a general criterion, I visually inspected each phylogenetic tree and selected hits retrieved from the MATOU-v2 dataset depending on their proximity to known diatom and non-diatom sequences and on bootstrap values of the clades they fell in. Unfortunately, due to technical reasons, the phylogenetic tree corresponding to

the ISIP2 sequences could not be obtained, and will thus not be discussed from now on. A brief description of each phylogenetic tree is found below.

NRAMP

NRAMP phylogenetic tree (Fig. 3.17) was made of 586 sequences out of which 532 were retrieved from MATOU-v2 and filtered according to bit score, taxonomic annotation and geographical distribution.

The tree was composed by a strongly supported clade where MATOU-v2 sequences clustered together with the *P. multistriata* query sequences. A second well supported clade (bootstrap value: 98) held the majority of diatom sequences, including the ones of raphid (*F. cylindrus*, *S. robusta*) and araphid (*S. acus*) Pennates and Centrics (*T. pseudonana* and *T. oceanica*); a high number of MATOU-v2 sequences fell in these clades, all selected for following analysis. Strangely, one *Emiliania huxleyi* sequence fell in this clade too; the MATOU-v2 sequences falling in the *E. huxleyi* clade were considered as non-diatoms.

The diatom-dominated cluster was separated from a strongly supported big clade (bootstrap value: 100) that held the majority of MATOU-v2 sequences and included a variety of non-diatom species ranging from the phylum Haptophyta, like the coccolitophore *E. huxleyi*, to Ocrophyta, represented by the class Pelagophyceae and Phaeophyceae (*Ectocarpus siliculosis*, considered as sister group of diatoms), as well as Chloropyta, represented by the *Ostrococcus* species *O. lucimarinus* and *O. tauri*. This clade also included sequences belonging to the terrestrial angiosperm *Arabidopsis thaliana* and contained one prokaryotic gene belonging to the extremophilic bacterium *Deinococcus radiodurans*. All the MATOU-v2 sequences falling in this clade were not included in subsequent analysis. Another small clade containing the polar centric diatom *C. cryptica* sequences seemed separated from the others. Strangely, *F. cylindrus* sequences did not cluster with any MATOU-v2 hit.

A total of 235 MATOU-v2 sequences were selected based on tree topology, out of which 223 were found only in surface samples and thus kept for the subsequent analysis.

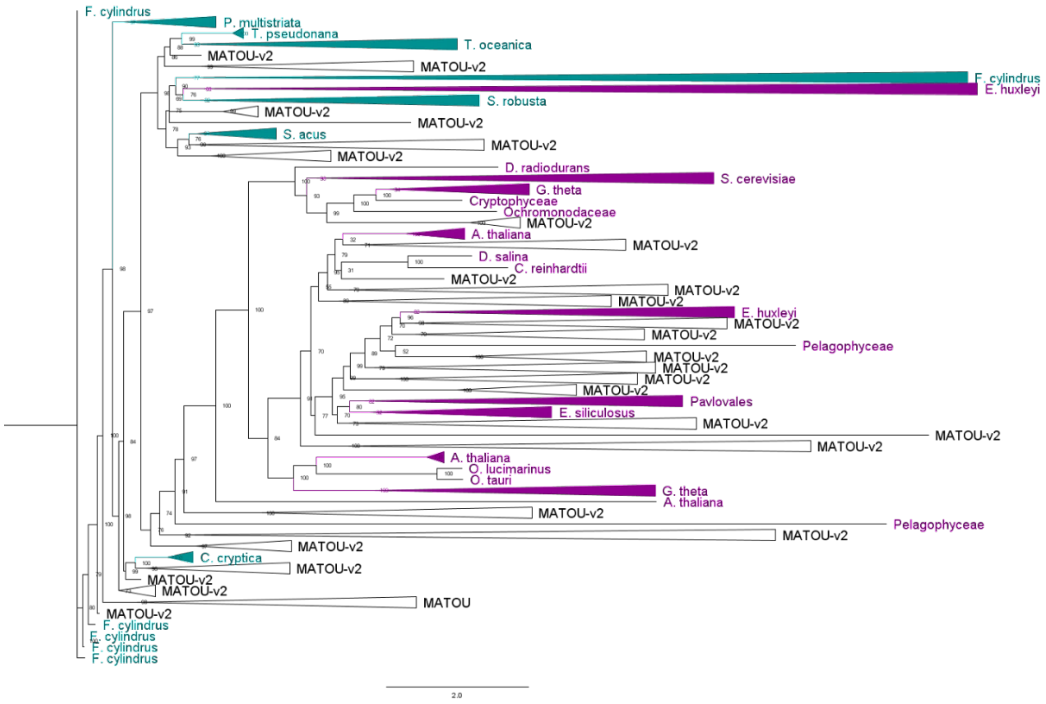

Figure 3.17. Maximum Likelihood (ML) tree of NRAMP genes.

Numbers at the base of each node refer to bootstrap support after 1000 replicates. Colours refer to the origin of sequences: in black sequences retrieved from the MATOU-v2 database, in purple non-diatom sequences, in pine green diatom sequences. The triangles indicate collapsed clades.

ZIP

ZIP phylogenetic tree (Fig. 3.18) was made of 695 sequences out of which 529 were retrieved from MATOU-v2 and filtered according to bit score, taxonomic annotation and geographical distribution. The tree covered a wide variety of organisms and it represented a broad exploration as it included genes from ZIP1 to ZIP16; in fact, it has to be noticed that the specificity of ZIP subclasses for iron transport is not confirmed.

A clade of *Pseudo-nitzschia* spp., holding both query sequences of *P. multistriata* and *P. multiseries*, was found next to another well supported clade of MATOU-v2

sequences, all selected for the subsequent analysis. Another diatom clade included queries of the raphid pennate *P. tricornutum* and *F. solaris*, while a separate group was mainly represented by the polar centric *T. oceanica*. Besides these diatom-specific clusters, other diatom sequences also fell in clades with other organisms. The lack of a unique diatom group cannot be explained by the use of different ZIP genes, that, in fact, still tended to group together more according to taxonomy than ZIP type (data not shown). Diatom sequences tended to form clusters per se but it was not always the case and there was not a clear separation between diatoms and non-diatoms. As a rule, for this marker gene I only retrieved sequences that clustered within diatom sequences with a bootstrap support of at least 60. The tree showed a low overall bootstrap support over many branches.

A total of 98 MATOU-v2 sequences were selected based on tree topology, out of which 96 were found only in surface samples and thus kept for the subsequent analysis.

Figure 3.18. Maximum Likelihood (ML) tree of ZIP genes. Numbers at the base of each node refer to bootstrap support after 1000 replicates. Colours refer to the origin of sequences: in black sequences retrieved from the MATOU-v2 database, in purple non-diatom sequences, in pine green diatom sequences. The triangles indicate collapsed clades.

ISIP1

ISIP1 phylogenetic tree (Fig. 3.19) was made of 909 sequences out of which 893 were retrieved from MATOU-v2 and filtered according to bit score, taxonomic annotation and geographical distribution.

Being first discovered and described in diatoms, ISIP genes are mainly studied in diatoms; therefore, I could only include one non-diatom sequence in the tree, i.e., the one from the green microalgae *Dunaliella salina*, that did not cluster with any of the MATOU-v2 sequences. Different ISIP1 genes from the same diatom species were found in different clades, like in the cases of *P. multistriata*, *F. cylindrus* and *T. oceanica*. A high number of MATOU-v2 sequences clustered together in unresolved clades that were clearly distant from the ones holding diatom sequences, and were thus discarded from the subsequent analysis.

A total of 229 MATOU-v2 sequences were selected based on tree topology, out of which 213 were found only in surface samples and thus kept for the subsequent analysis.

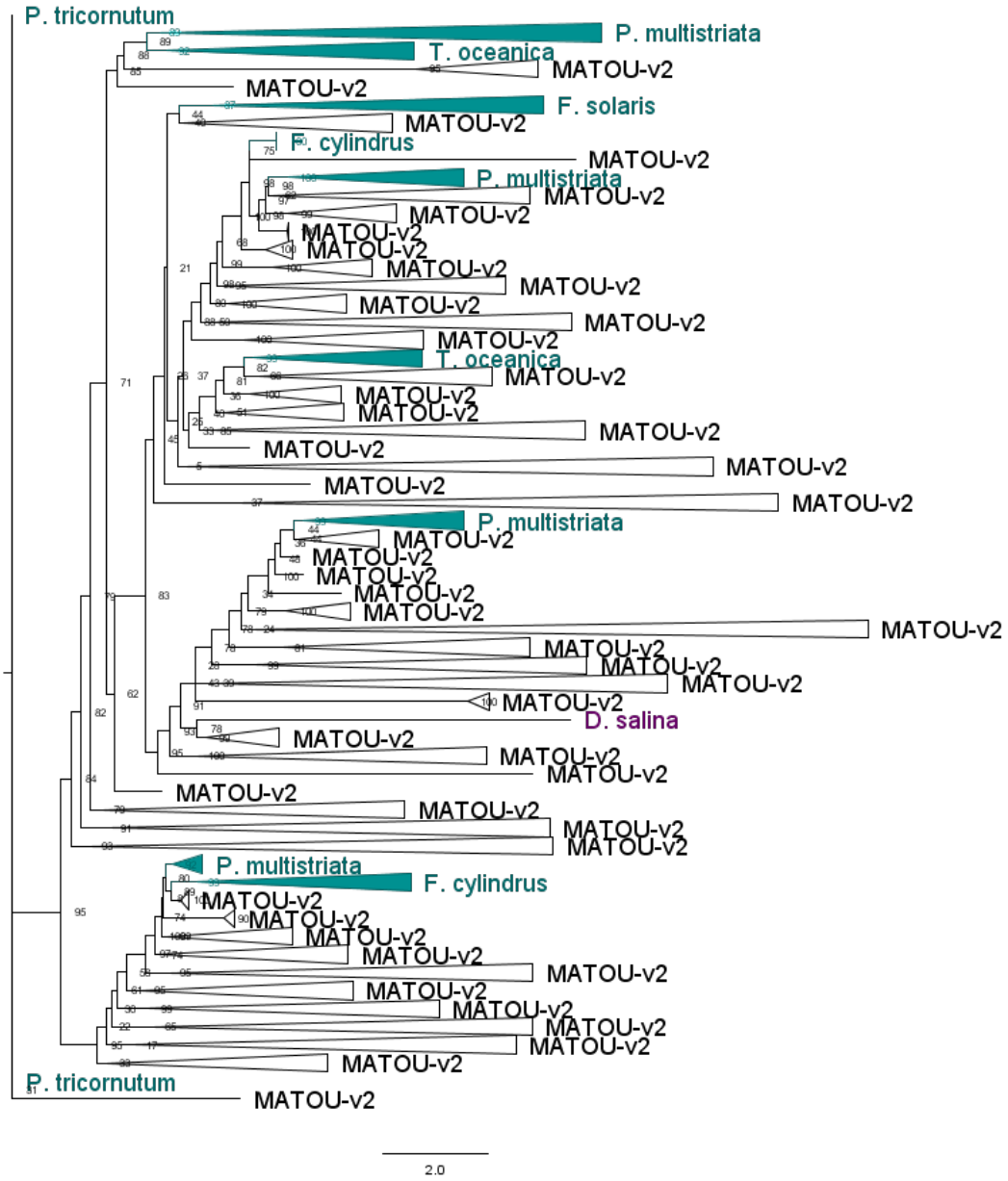

Figure 3.19. Maximum Likelihood (ML) tree of ISIP1 genes.

Numbers at the base of each node refer to bootstrap support after 1000 replicates. Colours refer to the origin of sequences: in black sequences retrieved from the MATOU-v2 database, in purple non-diatom sequences, in pine green diatom sequences. The triangles indicate collapsed clades.

FBP

FBP phylogenetic tree (Fig. 3.20) was made of 616 sequences out of which 573 were retrieved from MATOU-v2 and filtered according to bit score, taxonomic annotation and geographical distribution. The phylogenetic tree was made only by

diatom sequences since both orthologous gene families retrieved via Blastp search in the Plaza diatom database were diatom-specific. Several significantly supported clades were composed by MATOU-v2 hits and known diatom genes, like the ones holding sequences of *F. solaris*, *F. cylindrus*, *P. multistriata*, *T. oceanica* and *S. robusta*. Poorly supported branches with several MATOU-v2 sequences in absence of a diatom sequence were found, and thus excluded from the subsequent analysis.

A total of 171 MATOU-v2 sequences were selected based on tree topology, out of which 166 were found only in surface samples and thus kept for the subsequent analysis.

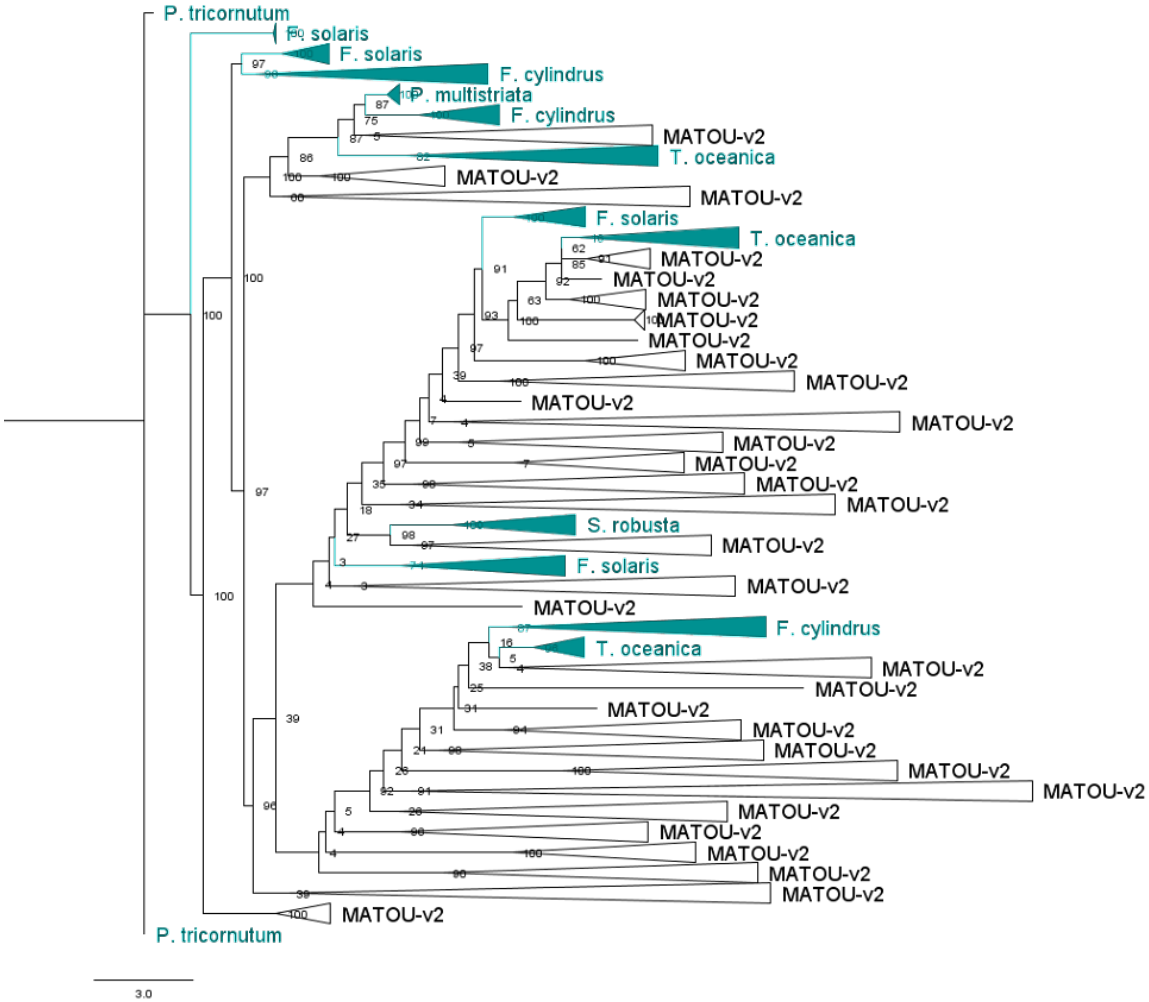

Figure 3.20. Maximum Likelihood (ML) tree of FBP genes.

Numbers at the base of each node refer to bootstrap support after 1000 replicates. Colours refer to the origin of sequences: in black sequences retrieved from the MATOU-v2 database, in purple non-diatom sequences, in pine green diatom sequences. The triangles indicate collapsed clades.

FRE

FRE phylogenetic tree (Fig. 3.21) was made of 750 sequences out of which 669 were retrieved from MATOU-v2 and filtered according to bit score, taxonomic annotation and geographical distribution.

The phylogenetic tree shows two distinct diatom clades that held sequences of many species without a clear clustering based on taxonomy. All sequences belonging to these two clusters were selected for the subsequent analysis.

A total of 212 MATOU-v2 sequences were selected based on tree topology, out of which 204 were found only in surface samples and thus kept for the subsequent analysis.

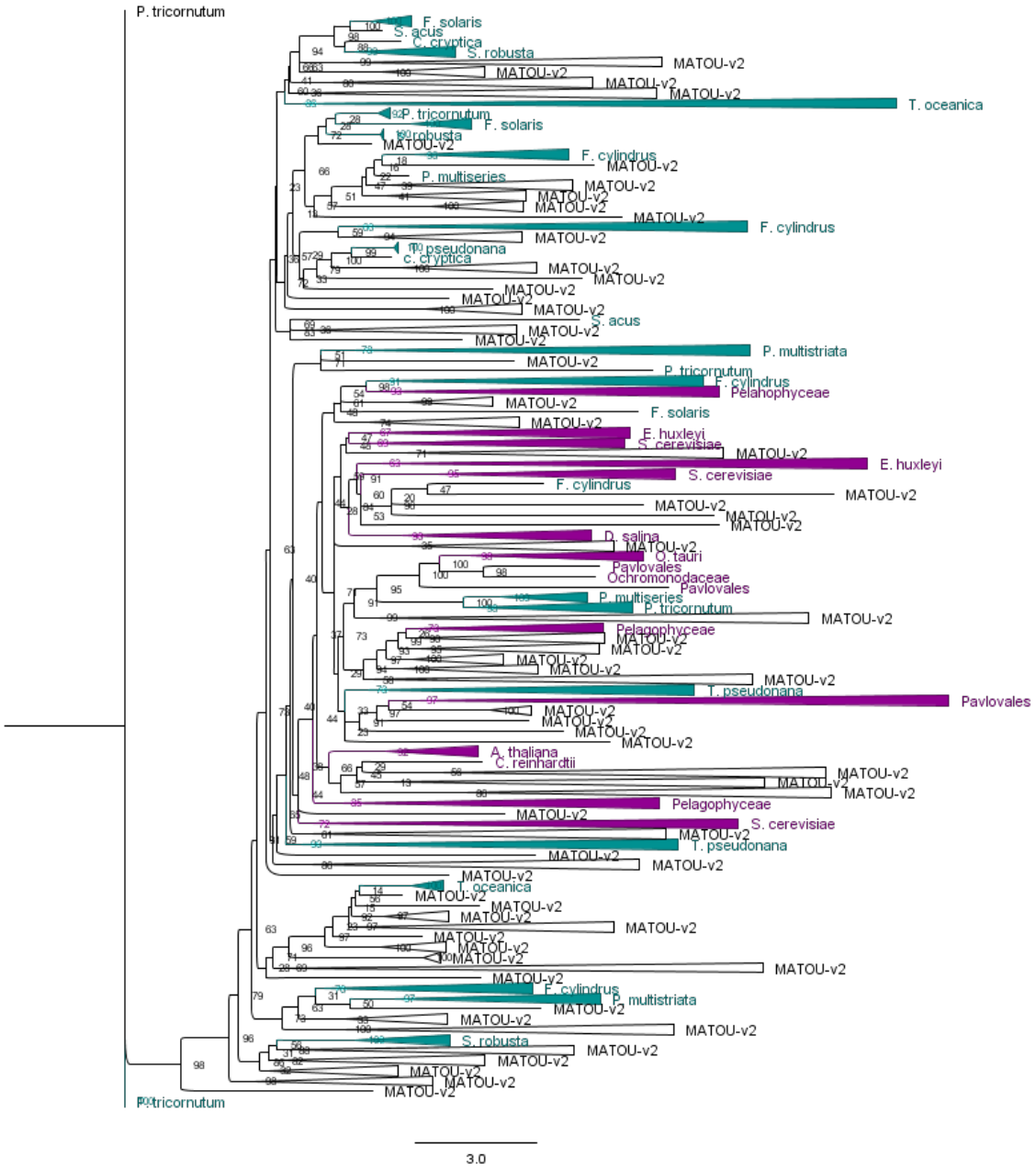

Figure 3.21. Maximum Likelihood (ML) tree of FRE genes.

Numbers at the base of each node refer to bootstrap support after 1000 replicates. Colours refer to the origin of sequences: in black sequences retrieved from the MATOU-v2 database, in purple non-diatom sequences, in pine green diatom sequences. The triangles indicate collapsed clades.

FET/MCO

FET/MCO phylogenetic tree (Fig. 3.22) was made of 1079 sequences out of which 997 were retrieved from MATOU-v2 and filtered according to bit score, taxonomic annotation and geographical distribution.

Bootstrap values were overall weak. A clade containing only diatom sequences, with a bootstrap support of 47%, held sequences from *S. robusta*, *T. oceanica*, *C. cryptica* and *F. cylindrus*. Only MATOU-v2 sequences falling in this clade with nodes showing a bootstrap support equal or higher than 60 were considered for the subsequent analysis. A strongly supported clade contained only non-diatom sequences, i.e., sequences belonging to *E. huxleyi*, *A. anophagefferens*, *C. reinhardtii*, *C. elegans*, *S. cerevisiae*; no MATOU-v2 sequences falling in this clade were included in subsequent analysis. A mixed clade of both diatom sequences like the ones of S. acus and *P. multistriata* and non-diatom sequences like the ones of *S. cerevisiae* or *C. reinhardtii* showed low bootstrap values; in this case only sequences strictly falling in the same clade of diatoms were selected. A relevant number of MATOU-v2 sequences remained unresolved by the phylogenetic analysis.

A total of 71 MATOU-v2 sequences were selected based on tree topology, out of which 69 were found only in surface samples and thus kept for the subsequent analysis.

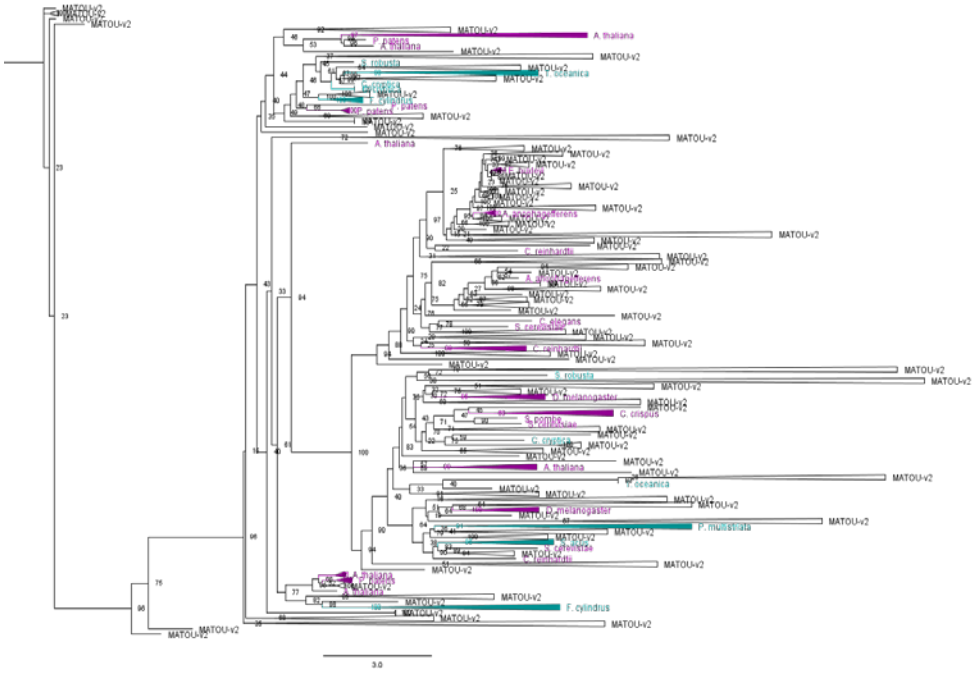

Figure 3.22. Maximum Likelihood (ML) tree of FET/MCO genes.

Numbers at the base of each node refer to bootstrap support after 1000 replicates. Colours refer to the origin of sequences: in black sequences retrieved from the MATOU-v2 database, in purple non-diatom sequences, in pine green diatom sequences. The triangles indicate collapsed clades.

FTR

FTR phylogenetic tree (Fig. 3.23) was made of 1079 sequences out of which 997 were retrieved from MATOU-v2 and filtered according to bit score, taxonomic annotation and geographical distribution.

The phylogenetic tree held a clear diatom-specific clade with sequences of the raphid pennates *F. cylindrus*, P. *multiseries*, *S. robusta* and *F. solaris*. All MATOU-v2 sequences falling in this clade were selected. Diatom sequences were also found in other clades, like the case of sequences of *T. pseudonana*, *P. multistriata* and *S. acus*. In these cases, only sequences strictly falling within the diatom clade were retained for subsequent analysis.

A total of 77 MATOU-v2 sequences were selected based on tree topology, out of which 73 were found only in surface samples and thus kept for the subsequent analysis.

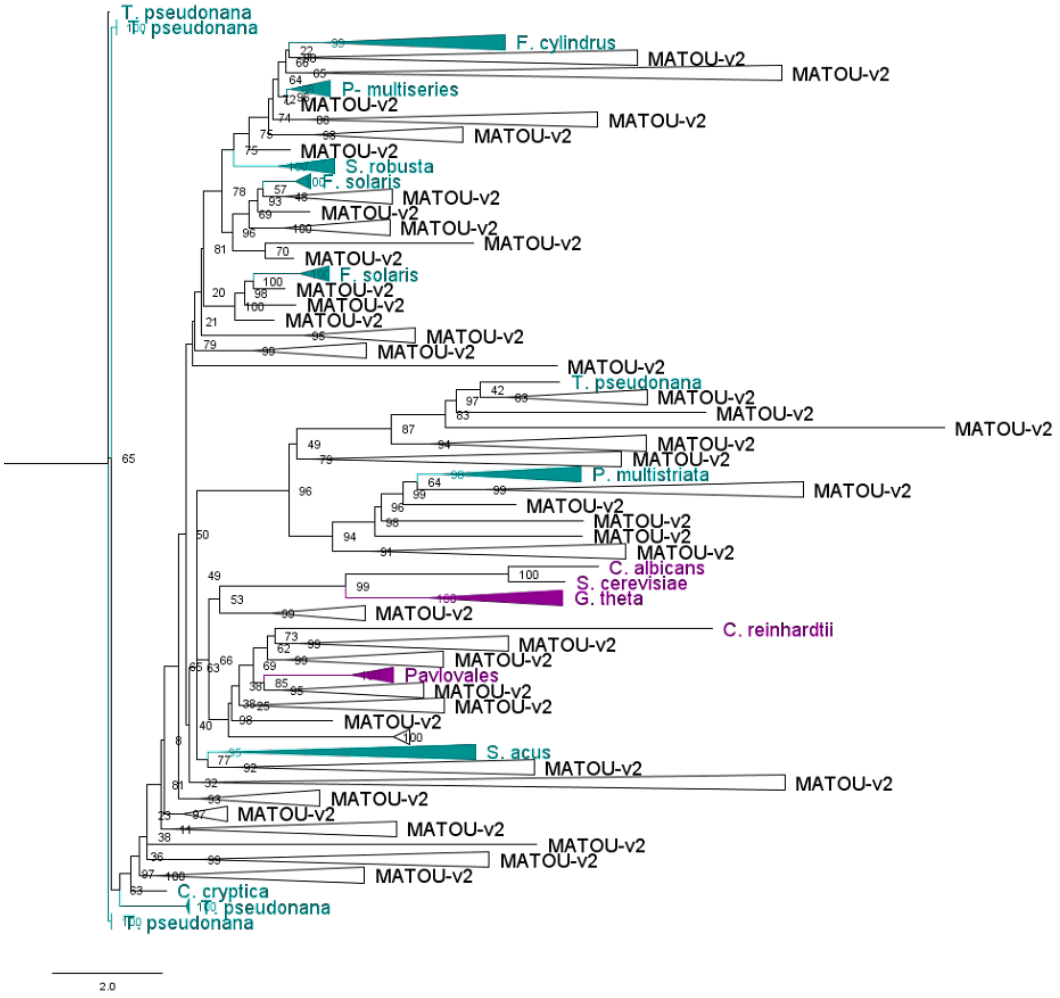

Figure 3.23. Maximum Likelihood (ML) tree of FTR genes.

Numbers at the base of each node refer to bootstrap support after 1000 replicates. Colours refer to the origin of sequences: in black sequences retrieved from the MATOU-v2 database, in purple non-diatom sequences, in pine green diatom sequences. The triangles indicate collapsed clades.

FTN

FTN phylogenetic tree (Fig. 3.24) was made of 215 sequences out of which 189 were retrieved from MATOU-v2 and filtered according to bit score, taxonomic annotation and geographical distribution.

The phylogenetic tree held one strongly supported diatom clade with the *P. multiseries* query from which all the MATOU-v2 sequences were selected. Another diatom clade was represented by sequences clustered with *P. tricornutum* query sequences. Non-diatom sequences formed a separate clade that includes *A.*

thaliana, *C. reinhardtii*, *D. salina*, *Ostreococcus* spp. (*O. tauri* and *O. lucirmarinus*), *G. theta* and other Cryptophyceae.

A total of 71 MATOU-v2 sequences were selected based on tree topology, out of which 67 were found only in surface samples and thus kept for the subsequent analysis.

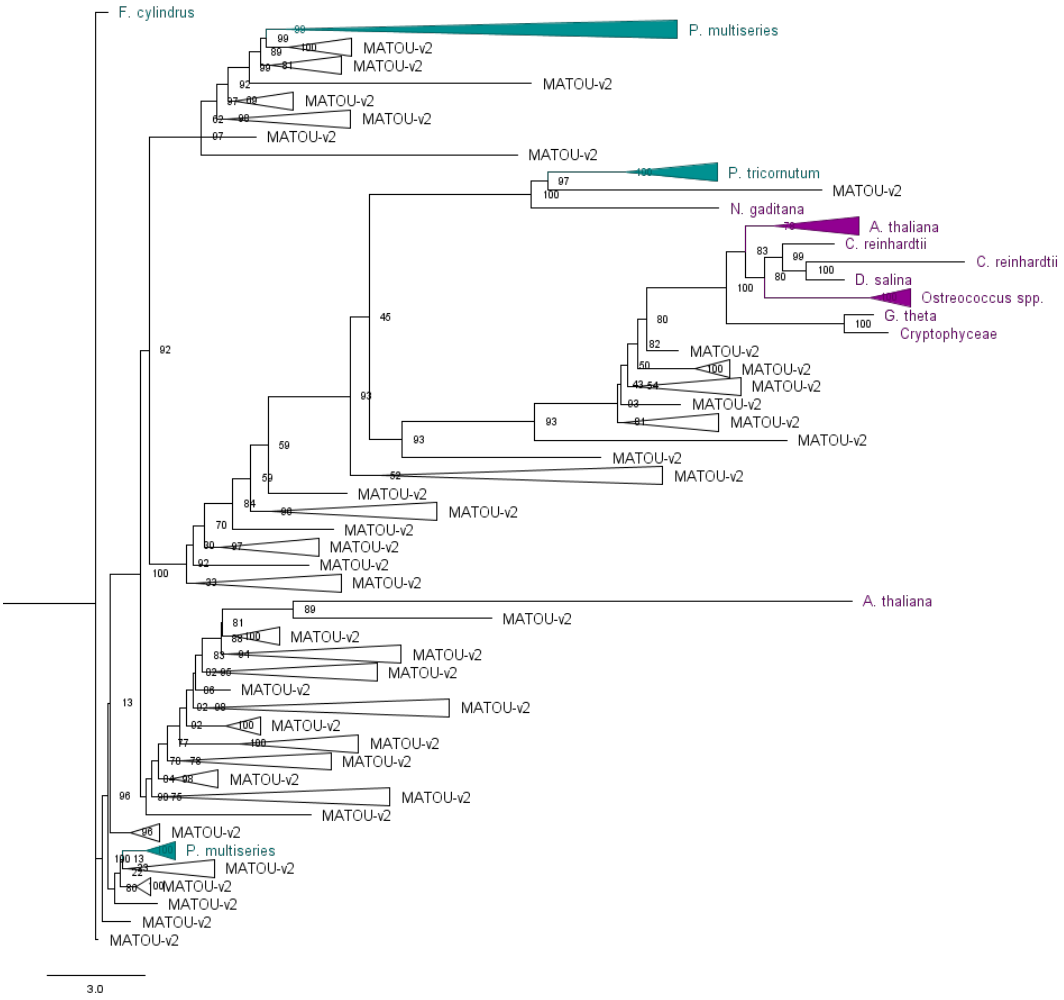

Figure 3.24 Maximum Likelihood (ML) tree of FTN genes.

Numbers at the base of each node refer to bootstrap support after 1000 replicates. Colours refer to the origin of sequences: in black sequences retrieved from the MATOU-v2 database, in purple non-diatom sequences, in pine green diatom sequences. The triangles indicate collapsed clades.

ISIP3

ISIP3 phylogenetic tree (Fig. 3.25) was made of 1826 sequences out of which 1807 were retrieved from MATOU-v2 and filtered according to bit score, taxonomic annotation and geographical distribution.

All MATOU-v2 sequences falling in the clade that held sequences of *Thalassiosira* spp. were kept. Other diatom sequences, from *F. cylindrus*, *F. solaris* and *P. tricornutum*, fell in other clades, while a clear non-diatom clade was composed by sequences of *G. theta*, other Cryptophyceae, as well as members of Pavlovales and Pelagophyceae. The overall low bootstrap support values did not allow a clear resolution of the taxonomic and functional assignation of MATOU-v2 hits. Therefore, although the numerous starting sequences, a total of 27 MATOU-v2 sequences were selected based on tree topology, out of which 26 were found only in surface samples and thus kept for the subsequent analysis.

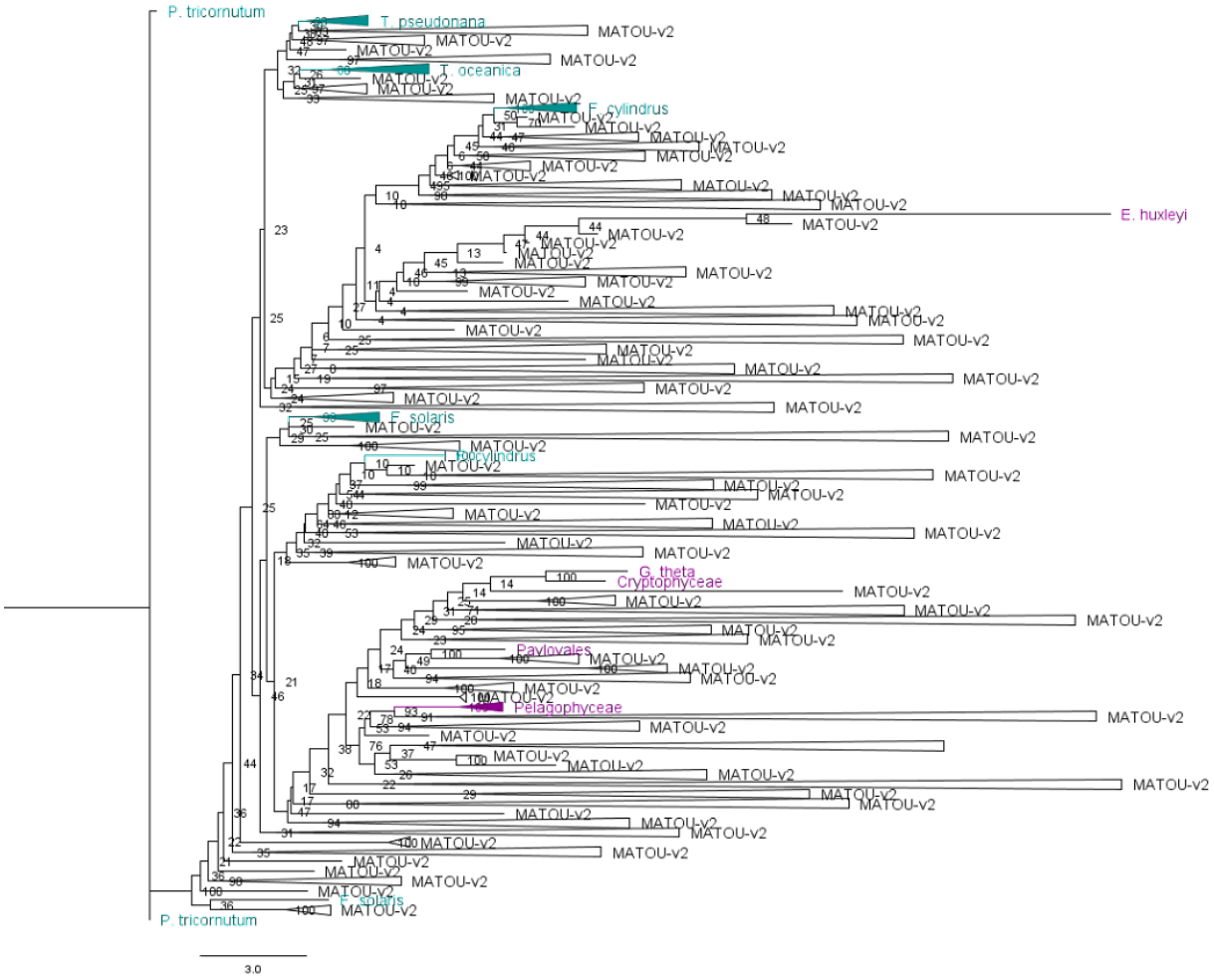

Figure 3.25. Maximum Likelihood (ML) tree of ISIP3 genes.

Numbers at the base of each node refer to bootstrap support after 1000 replicates. Colours refer to the origin of sequences: in black sequences retrieved from the MATOU-v2 database, in purple non-diatom sequences, in pine green diatom sequences. The triangles indicate collapsed clades.

Final set of genes

The final number of genes, retained after selection according to bit score values, taxonomy, geographical distribution and phylogenetic analysis is shown in Table 3.6.

Table 3.6. Final number of genes used.

Gene	Tot
NRAMP	223
ZIP	96
ISIP1	213
FBP	166
FRE	204
FET/MCO	69
FTR	73
ISIP3	26
FTN	67

3.3.3.3 Occurrence of markers across stations

The heatmap in Fig. 3.26 shows the result of Ward D2 clustering of stations and genes based on the Jaccard similarity indices calculated on the occurrence matrix.

The Figure displays a reconciliation between the patterns observed with environmental and metabarcoding data. When looking at the broad station clustering (rows), the North Atlantic and the Arctic Ocean appeared as two different systems, each holding its specific set of genes and thus species. However, when integrating the information from gene clusters (columns), a more complex scenario

is displayed. First of all, a high number of genes were held by station 155; although geographically belonging to the North Atlantic Ocean, this station lied in the same cluster of Arctic Ocean stations. Besides its similarity with the geographically close station 158, station 155 was composed by a high number of genes that were all unique to this station, the majority of them corresponding to ISIP1 and FBP, two proteins involved in the same mechanism of iron metabolism, i.e., the siderophore-mediated uptake of ferric ion (see Table 3.2). Besides this peculiar case, three main clusters of genes were shown, with a clear biogeographical signal. First, a polar cluster was characterized by genes exclusive to the Arctic Ocean. A subtropical cluster held subtropical genes that appeared in North Atlantic stations. A third cluster, with a subpolar/polar distribution, was held by stations belonging to the transition zone as well as the Arctic stations more influenced by waters influxes coming from both North Atlantic and Pacific Oceans.

An exception to this configuration was represented by the North Atlantic station 145, whose peculiarity was already clear from environmental and metabarcoding analysis (see Fig. 3.6A and Fig. 3.11). In particular, the environmental analysis showed that station 145 represents waters actually colder than the ones surrounding its geographically closer stations; the metabarcoding analysis indicated that diatom OTUs inhabiting this station are actually more similar to the ones found in the Arctic than in the North Atlantic Ocean, suggesting that the cold waters that flow in this station from northern regions bring also a relevant amount of OTUs. The results presented in the heatmap, obtained with metatranscriptomic data, are even more interesting since they confirm that species that flowed away from the northern Labrador Sea and went back into the North Atlantic Ocean were still metabolically active.

Both the North Atlantic and the Arctic groups were in turn composed of several sub clusters. The intra-oceanic clustering reflected the internal regionalization of the Arctic Ocean, showing three main groups: a set of stations influenced by the North Atlantic Ocean, a group more internal to the Arctic, and a cluster of stations

influenced by the Pacific Ocean. The two stations representing the transition between the North Atlantic and the Arctic Ocean, i.e., 155 and 158, clustered together and separately from the rest of the polar stations; while being upstream the Arctic, they shared a majority of their genes with arctic stations, further suggesting the existence of a subcommunity shared between subpolar and polar regions. The internal grouping of the North Atlantic Ocean also broadly followed the geographical distribution of stations. When looking at the gene composition of clusters, no clear patterns emerged from the analysis, suggesting that both the North Atlantic and the Arctic diatom communities were provided with the complete molecular toolkit to deal with iron uptake, transport and storage.

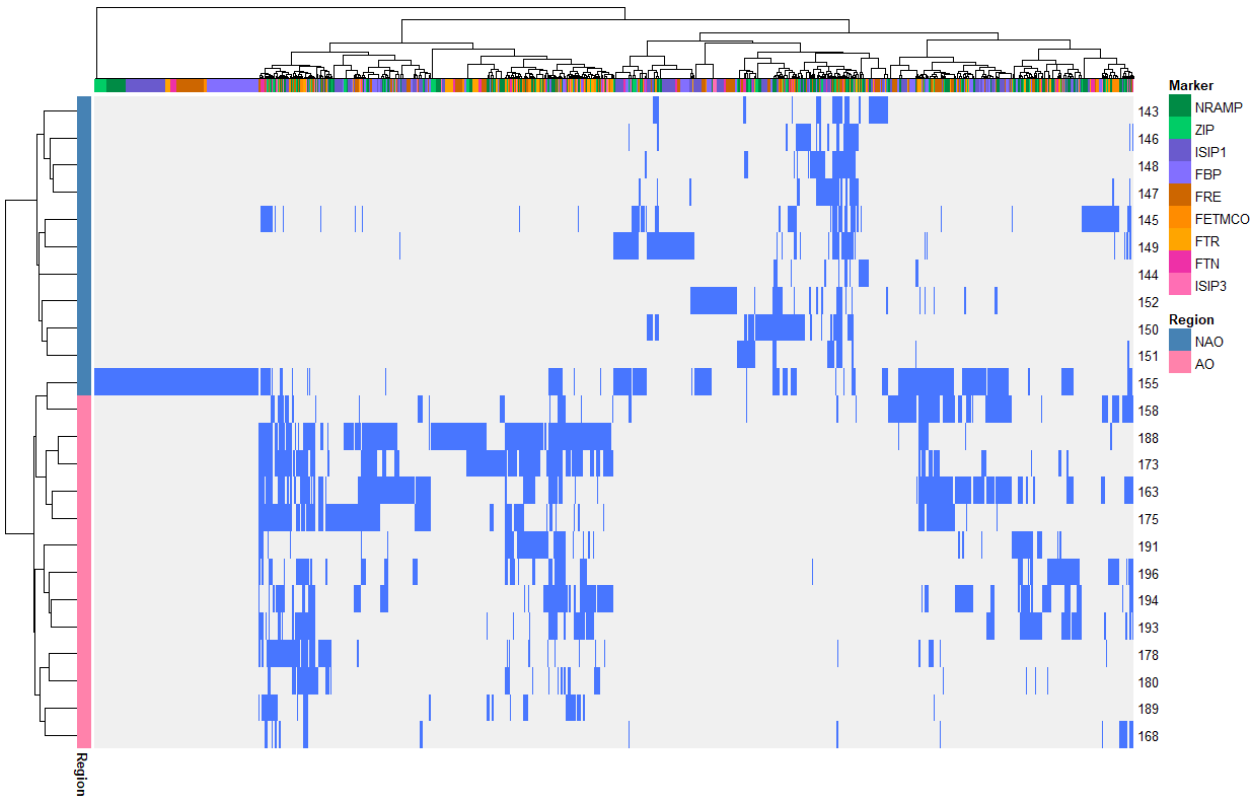

Figure 3.26. Heatmap showing Ward D2 clustering of stations (rows) and genes (columns) according to the Jaccard similarity indices calculated on the presence-absence matrix. The presence of a transcript in a station is indicated in blue, while its absence is coloured in grey. Station numbers are indicated and rows are coloured according to the geographic basin; colours of transcripts are based on the marker they belong.

3.3.3.4 Distribution of transcripts across diatom taxa

A high fraction of transcripts was not taxonomically assigned at the class level. This percentage of unknown taxa reached the 60% of transcripts in the case of FTN, and was on average the 30% of the retrieved sequences (Fig. 3.27A).

When the class annotation was present, I divided diatoms into Centrics (Coscinodiscophyceae and Mediophyceae) and Pennates (Fragilariophyceae and Bacillariophyceae). This subdivision allowed to have a first glimpse of differences between the two large groups of diatoms, differences that have been proposed but never confirmed for iron metabolism. As a general trend, centric diatoms showed a higher percentage of transcripts for ISIP1, FBP, FETMCO and ISIP3 genes and lower percentages for FRE compared to Pennates; moreover, they were almost absent in FTR. On the other hand, FTR transcripts belonged almost only to Pennates, that also expressed a high fraction of FRE genes, which is interesting since both genes are involved in the same iron uptake reductive process. FBP genes assigned to pennate diatoms were less than 10% of total.

For each gene I looked at the taxonomic assignation, when present, at genus level (Fig. 3.27B and C). *Pseudo-nitzschia* and *Fragilariopsis* were the most detected genera in pennate diatoms, while *Thalassiosira* dominated the Centric group. Almost half of ISIP1 transcripts did not hold any taxonomic assignation at the Pennate/Centric level; pennate diatoms expressing this gene belonged to *Pseudo-nitzschia* and *Fragilariopsis*, while centric diatoms were mainly assigned to *Thalassiosira* species, followed by *Proboscia* and *Chaetoceros*. FBP genes, whose distribution was correlated with the one of ISIP1 in a recent comparative transcriptomic study (*Gao et al., 2021*) showed indeed similar percentages of distribution among centric diatoms, while, among Pennates, it was almost only expressed by *Pseudo-nitzschia* species. FRE transcripts where predominantly assigned to pennate diatoms, and in particular to *Pseudo-nitzschia* and *Fragilariopsis* species. FET/MCO, an enzyme that uses Cu ions as cofactors to oxidize iron, was

instead mainly expressed by the centric *Thalassiosira*, a genus where iron uptake has been indeed shown to be dependent on copper (*Maldonaldo et al., 2006*).

FTR was almost only assigned to pennate diatoms, and especially to *Pseudo-nitzschia* spp. while, among Centrics, it was mainly expressed by *Thalassiosira*, *Skeletonema* and *Coscinodiscus* spp. Out of the annotated sequences, the majority of FTN genes assigned to Pennates belonged to *Pseudo-nitzschia* spp., a genus whose species thrive in iron-limited regions and for which this gene has been hypothesized as a fundamental storage protein (*Marchetti et al., 2009, 2017*). Moreover, studies suggested that this gene could play a role in *Pseudo-nitzschia* also in the cellular homeostasis of this micronutrient as a Fe^{2+} oxidase (*Pfaffen et al., 2013*), as it happens for higher plants, pointing to the multiple functional facets of this conserved gene (*Gao et al., 2021*). No FTN was found to be assigned to the centric *Thalassiosira*, a genus where iron storage is thought to occur with a different mechanism, i.e., storage in vacuoles (*Nuester et al., 2012*). In *Thalassiosira*, as well as in other (mainly centric) diatoms supposed to rely on vacuoles to deal with iron storage, the process involves the release of the micronutrient from vacuole, when needed, through divalent metal transporters like ZIP and NRAMP proteins; ZIP proteins are in my results almost all attributed to *Thalassiosira* when looking at Centrics. However, ZIP and NRAMP can also be part of Fe^{2+} uptake and are general divalent metal transporters not specifically addressed to iron. Finally, a putative role in storage has been suggested for ISIP3, since it holds a protein domain, DUF295, that belongs to the ferritin family (*Behnke and LaRoche, 2020*). In my analysis, ISIP3 held a lower percentage of non-assigned genes, and the majority of transcripts were ascribed to centric diatoms. At genus level, the only pennate diatom holding ISIP3 transcripts was *Fragilariopsis*, in agreement with the overabundance of this gene in this genus found in a recent study (*Gao et al., 2021*). ISIP3 transcripts were also abundant in the centric *Thalassiosira* and *Skeletonema*.

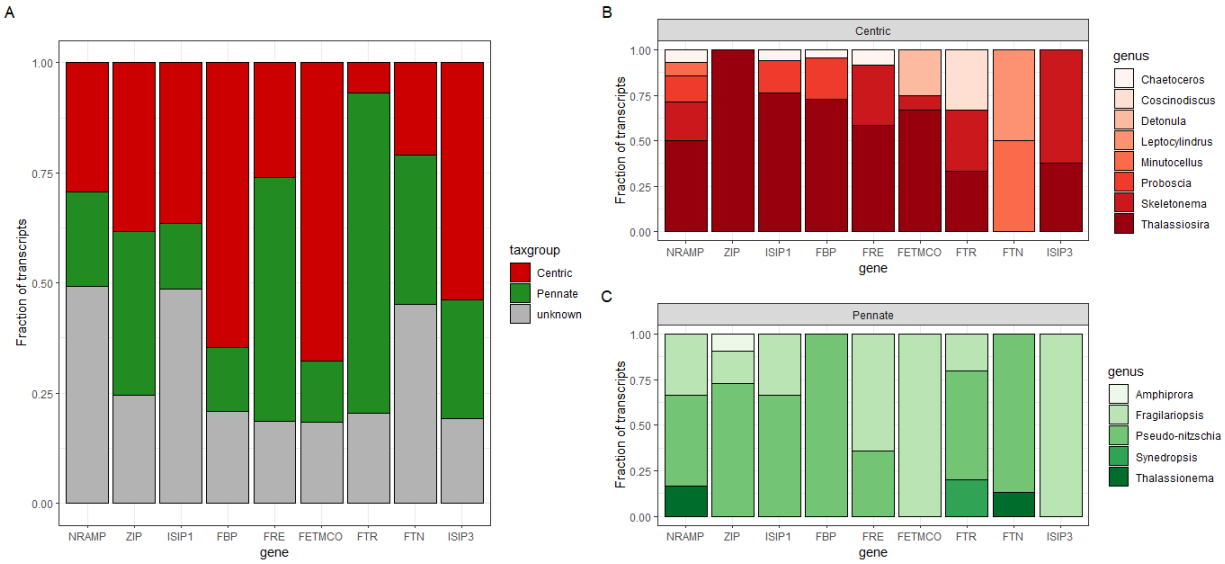

Figure 3.27. Bar plots showing the taxonomic assignation of transcripts at the A) class level and, when present, at the genus level for B) centric and C) pennate diatoms.

3.3.3.5 Gene expression from the North Atlantic to the Arctic Ocean

The geographical pattern observed when looking at the sole presence of genes (Fig. 3.26) becomes more complex with the integration of their relative expression information (Fig. 3.28). A high functional heterogeneity was observed, as both North Atlantic and Arctic Ocean expressed genes involved in different processes. As a broad observation, we could identify a higher expression of genes involved in the reductive iron uptake (i.e., FRE, FET/MCO, FTR) in Pennates compared to Centrics.

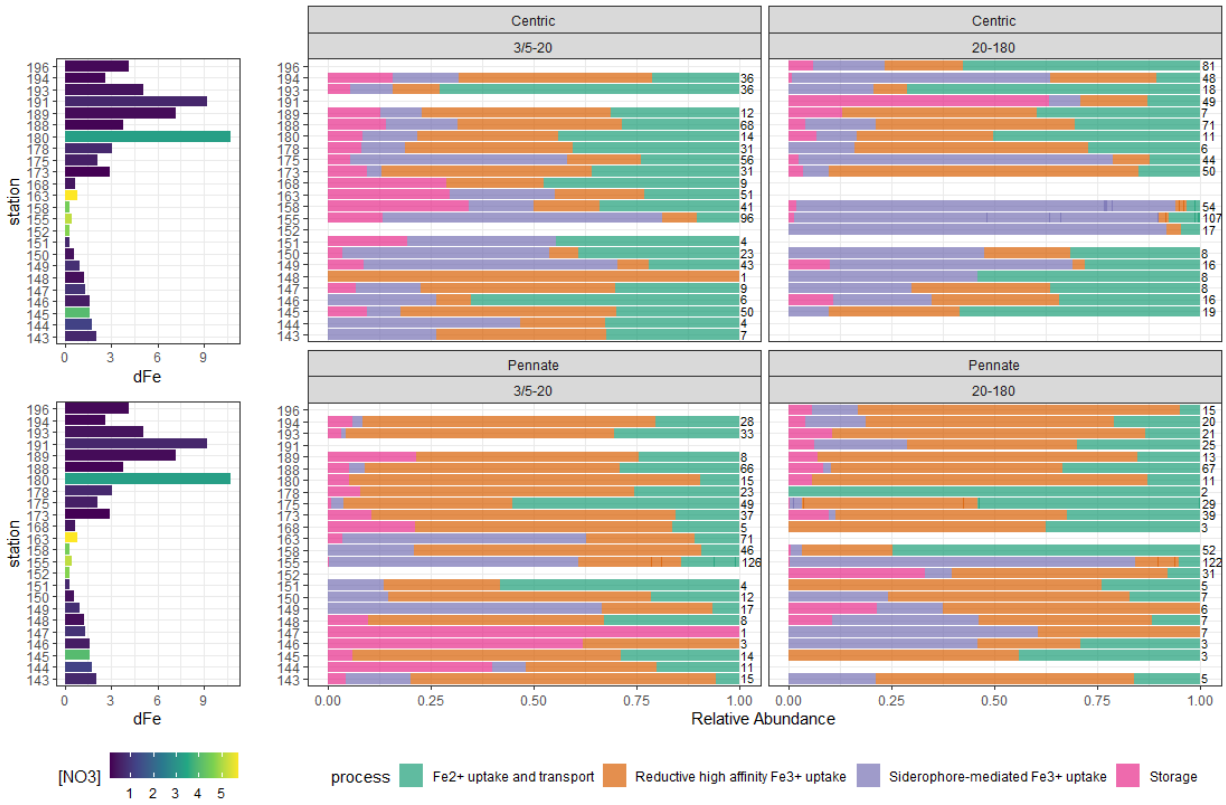

Figure 3.28. Bar plots showing the relative expression of transcripts belonging to the main processes involved in iron metabolism, according to their size class and taxonomic grouping. Numbers on the right side of each bar indicate the total richness for that sample and taxonomic group. A bar plot on the left shows in parallel the corresponding concentrations of iron and nitrate in each station.

The differences between the two main diatom groups became clearer when we looked at the abundances of single genes along the North Atlantic – Arctic Oceans (Fig. 3.29).

In centric diatoms, divalent metal transporters were mainly expressed through ZIP genes in the North Atlantic, while a higher amount of NRAMP genes was expressed in the Arctic Ocean. Siderophore-mediated iron uptake genes (ISIP1 and FBP) often co-occurred; the observed preponderance of these genes in the stations located at the transition between the two oceans was confirmed by the expression data; in particular, they constituted the most expressed genes for microplanktonic (size class 20-180 µm) centric diatoms. Interestingly, these genes, whose expression is enhanced by the iron starvation, showed the highest expression in stations where iron concentration was at its minimum. Genes involved in the reductive uptake of

iron (i.e., FRE, FET/MCO and FTR), supposed to act together, were often co-expressed, but a higher relative abundance of FET/MCO was found in the first stations of the North Atlantic, while FRE genes were highly expressed in the Arctic Ocean. Proteins involved in iron storage were overall low expressed in the North Atlantic than the Arctic. Pennate diatoms often showed opposite patterns compared to Centrics. For example, the two divalent metal transporter markers showed an inverted trend: NRAMP genes were mainly expressed in the North Atlantic Ocean, while ZIP transcripts were more abundant in the Arctic Ocean. A consistent pattern between the two main diatom taxa investigated is the one related to genes responsible for siderophore-mediated iron uptake; although less expressed in Pennates than in Centrics, both ISIP1 and FBP were particularly abundant in the North Atlantic – Arctic transition region. Storage proteins were overall low expressed.

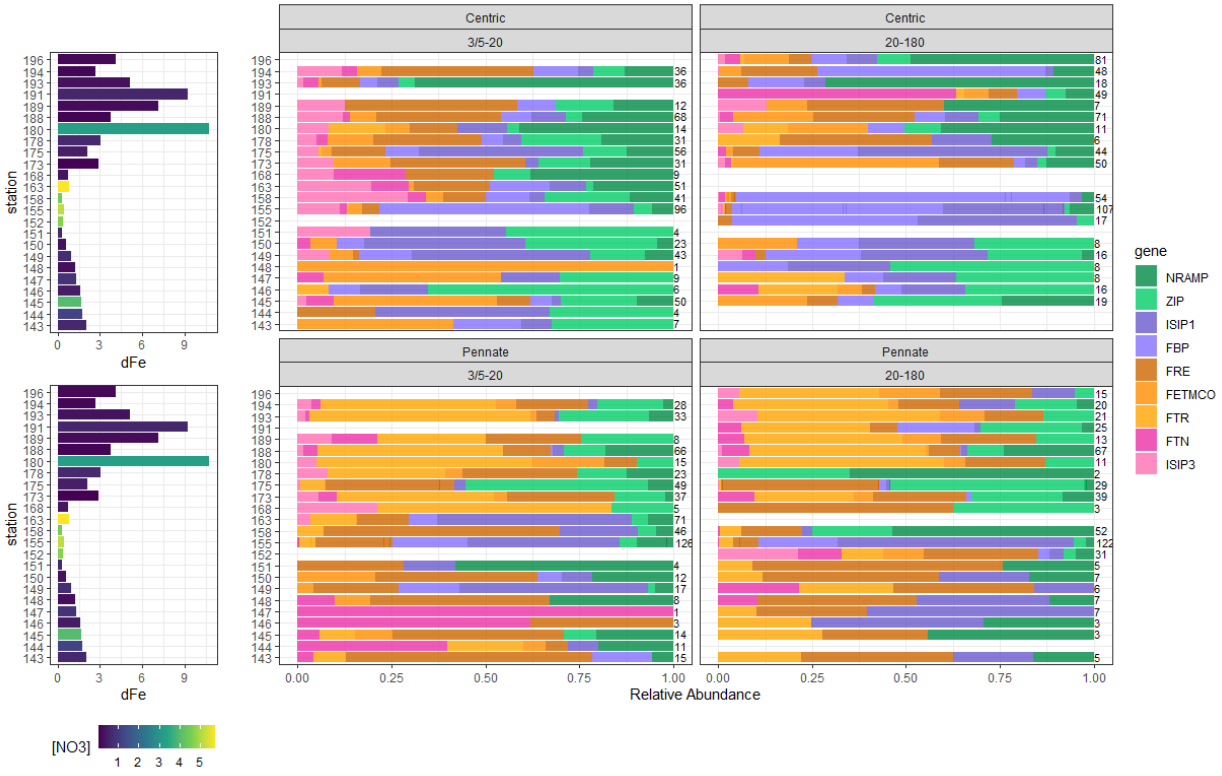

Figure 3.29. Bar plots showing the relative expression of transcripts belonging to the main genes involved in iron metabolism, according to their size class and taxonomic grouping. Numbers on the right side of each bar indicate the total richness for that sample and taxonomic group. A bar plot on the left shows in parallel the corresponding concentrations of iron and nitrate in each station.

The relative expression of genes would have been more informative if accompanied by the relative abundance of genes in metagenomic data. However, when selecting for each sample the same set of genes occurring in both data sets, a high quantity of information was lost (Fig. 3.30). Interestingly, the percentage of "transcript loss" was lower in the Arctic stations and higher in the North Atlantic, where it often exceeded half of the station richness, possibly due to the low diatom concentrations in that basin. For this reason, I decided to not include the comparison between metagenomic and metatranscriptomic abundance information.

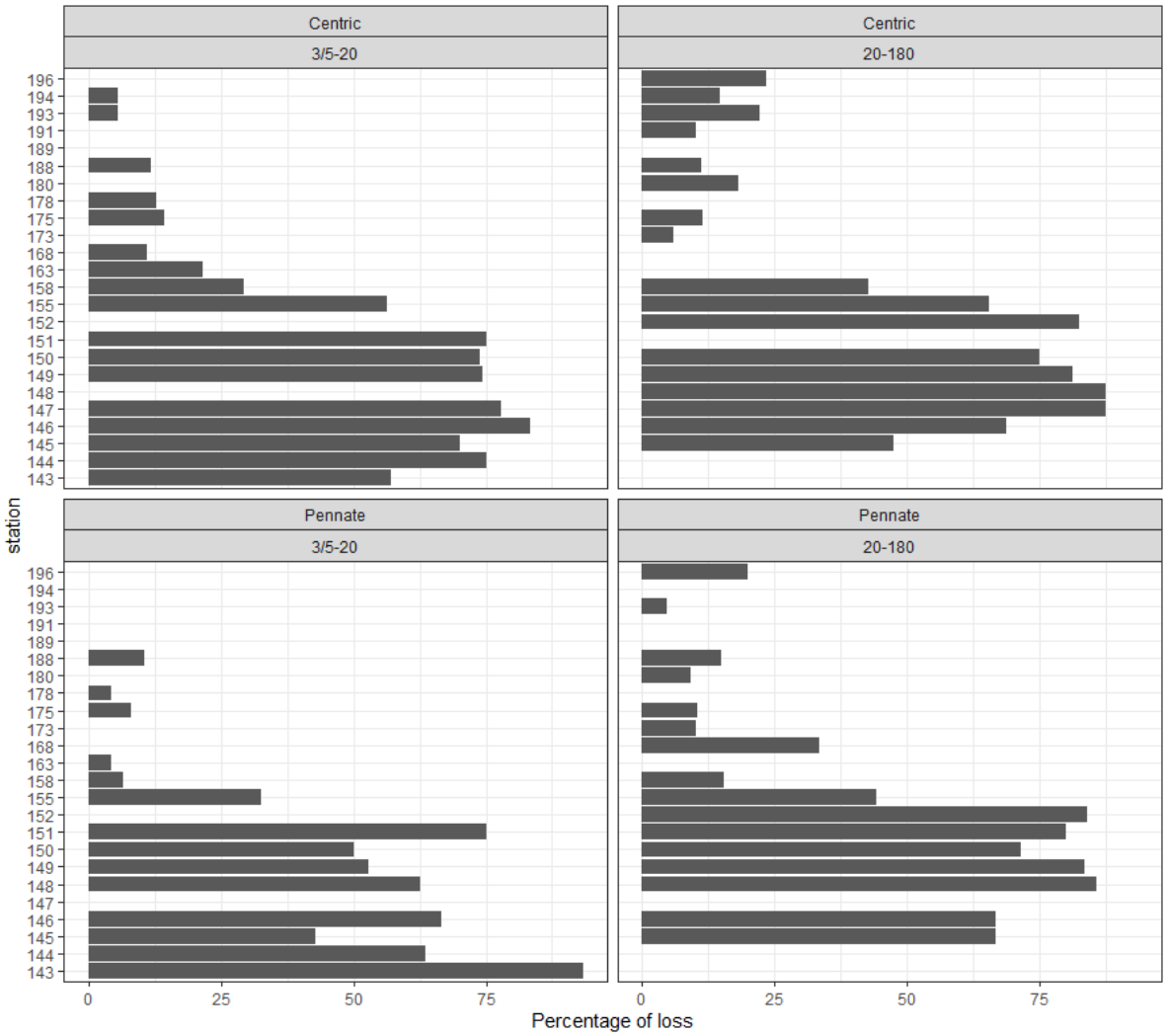

Figure 3.30. Bar plots showing the percentage of genes that are lost for each station, size and taxonomic group when metatranscriptomic data set is subset using the genes occurring in metagenomic data.

3.4 Conclusion

In this chapter I explored diatom communities along currents that bring Atlantic waters into the Arctic Ocean. Planktonic species are drifters immersed in a fluid that moves, transporting nutrients as well as organisms from a location to another in the ocean. Communities change across currents; some species survive and other undergo extinction. What we observe in a certain location and a certain time in the ocean is the consequence of multifaceted processes interacting together; seascape features, like horizontal advection and vertical mixing, together with chemical processes, light and nutrient availability, shape the biotic assemblages inhabiting the oceans. The data explored here, including environmental factors as well as diatom OTUs and functional transcripts, indicate how the North Atlantic and the Arctic Ocean are two separate systems. As shown in Chapter II, diatoms are abundant at poles, and an overall higher richness is found in the Arctic Ocean. However, beta-diversity analysis pointed to the role of species replacement more than richness difference as the main driver of the observed diversity between the two oceanic regions investigated in this chapter. The difference in the identity of the OTUs inhabiting the North Atlantic and the Arctic Ocean was reflected at the genus level. Three main communities of diatoms could be detected when exploring their taxonomy at this rank. In particular, sampling stations located at the interface between the two oceans, representing the gateway from North Atlantic to Arctic Ocean, showed a higher fraction of OTUs common to both basins, thus representing a real transition from a regime to another. These stations, that also held a certain amount of OTUs never found elsewhere, were taxonomically characterized by the dominance of two genera, namely *Pseudo-nitzschia* and *Fragilariopsis*, whose abundance was never high in the other sampling stations. This suggests that the transition zone is not simply a mixture of what is found in both North Atlantic and Arctic Oceans, but is characterized by its own diatom community, the exclusive subpolar community.

In summary, we observed a subtropical community, a subpolar community, a community that is made of subpolar and polar species and an Arctic exclusive community. Thus, a first message that emerges from the analysis I presented here is that what is found in Arctic Ocean is not a subset of what is flowing from the North Atlantic: some diatoms are equipped with the necessary molecular and physiological ability to survive the harsh polar environment, while others undergo local extinction. Moreover, the high richness found in the Arctic Ocean is an indication of a possible high speciation rate in this region, as already hypothesized (*Galhardo et al., 2007*; *Rabosky et al., 2018*). The environmental characterization of the Arctic stations confirmed the presence of high concentration of nutrients essential for diatoms while the subpolar region in Spring (i.e., at the moment of the sampling) was probably iron limited.

Other than confirming the separation between the two basins, the results obtained allowed the detection in the North Atlantic Ocean of diatom OTUs that flowed in the opposite way, i.e., transported from the Arctic Ocean back to the North Atlantic. The metatranscriptomic confirmation of the observed pattern, emphasising an even sharper separation between subtropical and subpolar/polar regions, is a strong indication that the species that were transported from the Labrador Sea back to the North Atlantic Ocean are also metabolically active and thus contributing to the functioning of the entire community. The implications of my results in the context of climate change are hard to predict: the two oceanic regions are separate systems, but the presence of a transition region often including stations that geographically belong to the Arctic Ocean probably reflects the ongoing change. This transition zone, including the western Arctic basin, is the region predicted to extend in the next years, where a shift from a cold stratified environment to a warmer and mixed "Atlantified" regime is supposed to occur (*Lind et al., 2018*). Addressing the consequence of this shift is still challenging, and the use of time-series data, and the seasonal information, would be indispensable for a more in depth understanding of the possible future scenarios.

Conclusion

The complex and stringent metatranscriptomic pipeline described in the second part of the chapter allowed the selection of genes whose taxonomical and functional assignation is robustly identified, although a variable number of transcripts did not show a corresponding taxonomical assignation. A first exploration based only on presence-absence data depicted a heterogeneous distribution of iron genes in both basins, and the pattern was confirmed by the abundance data. In particular, the comparison among genes through the computation of the ratio between functions allowed me to characterize how all the main genes involved in iron uptake, transport and storage are distributed relatively to each other across the oceanic currents. When looking at the process level, i.e., considering the uptake and transport of Fe^{2+}, the siderophore-mediated uptake of Fe^{3+}, its reductive uptake, and its storage, no clear biogeographical or taxonomical pattern emerged. However, when increasing the functional resolution of the investigation, I observed a regulation of iron metabolism that involved distinct markers in the two oceans; moreover, the two main groups of diatoms, i.e., Centrics and Pennates, showed different strategies.

The functional pipeline presented here can be applied to any functional process for which genomic or transcriptomic information is available in public databases. One future important step would be to extend the exercise presented here to other relevant metabolic genes, like the ones involved in nitrate, phosphate or silicate acquisition, with the aim of understanding how the main functional processes vary across currents.

Chapter IV

Metabarcoding to study *Pseudo-nitzschia* Biogeography and Ecology

4.1 Introduction

Environmental DNA (eDNA) is defined as any trace of DNA directly collected from environmental samples, like water, soil, or sediment, that represents a mixture of the genetic material belonging to many different organisms (*Bohmann et al., 2014*; *Thomsen and Willerslev, 2015*). Environmental DNA is extremely powerful in studying microbial diversity, since the majority of microorganisms are impossible to cultivate (*Brock, 1987*; *Pedrós-Alió, 2006*). The increasing efforts of eDNA sequencing over the last couple of decades have indeed allowed the identification of many novel species and taxa that have never been cultivated, and incredibly expanded our knowledge of microorganisms' biodiversity and distribution. The majority of eDNA production relies on the sequencing of the small subunit of ribosomal DNA, also called 18S rDNA; its presence in all eukaryotes, together with its structure made of conserved regions suitable for primer design and variable regions ideal for taxa identification makes this region the perfect eukaryotic "barcode", i.e. the best universal DNA marker available to date (*Pawlowski et al., 2012*), with the V4 and V9 regions of the 18S rDNA as the markers of choice in the majority of studies. The high throughput sequencing (HTS) of these taxonomically discriminative eDNA barcode regions, i.e., HTS metabarcoding, has substantially enhanced our ability to assess microorganisms' diversity from a mass collection of species (*Del Campo et al., 2018*, and references therein).

As shown by several projects, including Malaspina (*Duarte, 2015*), *Tara* Oceans (see paragraph 1.4.1.1), but also BioMarks (http://www.biomarks.eu), Ocean Sampling Day (OSD; https://www.microb3.eu/osd.html) and Long-Term Ecological Research Networks (LTERs; *Stern et al., 2018*), the use of HTS metabarcoding is a common practice in large spatial and temporal studies on marine microorganisms, including diatoms. In particular, DNA metabarcoding allowed the assessment of diatom biogeography and diversity from community (e.g., *Malviya et al., 2016*; *Piredda et al., 2018*; see paragraph 1.4.1.1.) to family level, like in the case of Leptocylindraceae (*Nanjappa et al., 2014*) to genus level, like for *Chaetoceros* spp. (*De Luca et al., 2019*), recently also providing interesting insights on the resolution of cryptic species within the *Chaetoceros curvisetus* species complex (*De Luca et al., 2021*). Metabarcoding also contributed to the assessment of the cosmopolitan distribution and worldwide abundance of *Pseudo-nitzschia* spp. (*Malviya et al., 2016*). Although important advancements in understanding *Pseudo-nitzschia* diversity and distribution (see also paragraph 1.3.1.), this ubiquitous, abundant and specious diatom genus has never been explored at global scale at high taxonomic resolution. Two main gaps are currently to be filled in *Pseudo-nitzschia* biogeography investigation; first, the majority of spatial studies of *Pseudo-nitzschia* species have been focused on temperate regions, thus overlooking both tropical waters as well as polar environments. Second, many sequences in GenBank have not been updated with species level annotation, and the need of curated references remains a major issue in *Pseudo-nitzschia* biogeographical and ecological studies (*Bates et al., 2018*).

The question of having a curated barcode reference library to interpret HTS data is indeed a common problem in metabarcoding studies. In principle, a reference database should be as comprehensive and as precise as possible, in order to allow a correct match of any new sequencing data with high quality and up-to-date taxonomically assigned reference sequences. Building such extensive and accurate databases requires careful cure by taxonomists and, in the last years, efforts have

been made to gather experts and create some universal resources. The two main examples for protists are SILVA (*Quast et al., 2012*), that also includes Bacteria and Archaea, and PR^2 (*Guillou et al., 2012*); moreover, taxon-specific databases have been developed, like EukRef-Ciliophora for ciliates (*Boscaro et al., 2018*) Dinoref for Dinophyceae (*Mordret et al., 2018*), Diat.barcode for diatoms (*Rimet et al., 2019*). SILVA and PR^2 databases will be further enriched by the contribution of Eukref, a community effort that, as part of the UniEuk project (unieuk.org), aims to complement existing databases with improved taxonomic information for eukaryotes (*del Campo et al., 2018*). Curated databases are crucial for the advancement of our knowledge of biodiversity: an exhaustive reference database of accurately taxonomically defined sequences would lead to a straightforward processing of metabarcoding data gathered from environmental samples from any location or time point directly into comprehensive list of species, even at lower taxonomic ranks like the genus and species levels (*Gaonkar et al., 2020*).

Species delimitation from cleaned HTS sequencing reads requires several steps, and taxonomic assignation comes only after another critical procedure in read processing, i.e., clustering or denoising, with the resulting construction of Operational Taxonomic Units (OTUs) or Amplicon Sequence Variants (ASVs), respectively. The common aim of these two processing steps is to avoid the impact of biases intrinsic to metabarcodes (e.g., the multicopy nature of ribosomal DNA) or to the laboratory procedures, like errors in DNA amplification and sequencing, and to reduce computational efforts by removing redundant diversity in the data (*Burki et al., 2021*). Both OTUs and ASVs are intended as biological units to which address downstream analysis on biodiversity and biogeography (*Callahan et al., 2017*; *Santoferrara et al., 2020*; *Forster et al., 2020*), but the degree of their reliability as comprehensive biological units is still debated, especially at low taxonomic ranks. For example, recent studies that measured within-OTUs variation with standard genetic approaches (*Elbrecht et al., 2018*; *Turon et al., 2020*) or with network analyses (*Forster et al., 2020*) demonstrated how OTUs are internally not homogeneous. It is

known that intraspecific variation may lead to speciation (*Coyne and Orr, 2004*) and it is therefore important to account for within OTUs variation when studying biodiversity and ecology of organisms. Moreover, the number of OTUs produced through metabarcoding processing strongly depends on the similarity cut-off thresholds used for clustering, and the lack of an "universal" threshold makes it difficult to find a trade-off between the "rare biodiversity" inflation provided by too high thresholds, and the underestimation of diversity due to too low thresholds (*Gaonkar et al., 2020*; *Burki et al., 2021*). Whether a threshold is too stringent or too relaxed strongly depends on the evolution rate of the target taxon, and the swarm approach (*Mahé et al., 2014*), used to build OTUs in the context of *Tara* Oceans, tried to overcome the problem of choosing a single clustering threshold by implementing an algorithm that uses a local threshold to cluster amplicons into OTUs. On the other hand, ASV approaches treat microvariations in HTS data as sequencing errors, thus ignoring by construction the minor haplotypes, with the risk of underestimating biodiversity, neglecting rare taxa and lumping together closely related species (*Gaonkar et al., 2020*). These problems are in part compensated by the highly reduced computation effort required by OTUs and ASVs analysis respect to what would occur with single ribotypes, and by the high amount of information that can still be retrieved from these processed data, especially at community level. Nevertheless, ideally, the use of ribotypes, i.e., unique sequences not processed to build OTUs or ASVs, would be the most accurate way of analysing metabarcoding data, and a recent study recommended this type of data for species-level biodiversity investigations in dinoflagellates (*Mordret et al., 2018*).

Species-level biodiversity is particularly important for diatoms belonging to the genus *Pseudo-nitzschia*. These raphid pennate diatoms, in fact, represent one of the most specious diatom genera, holding 58 species (*Dong et al., 2020*; *Guiry and Guiry, 2021*) out of which at least 26 are toxigenic (*Bates et al., 2018*), i.e., capable to cause Harmful Algal Blooms (HABs) through the production of the neurotoxin

domoic acid (DA; see section 1.3.1.). Toxicity is a critical aspect of this genus, that makes *Pseudo-nitzschia* spp. particularly important to study from both a bio-ecological and socio-economical point of view. In fact, it has been suggested that also species that are now labelled as harmless have the potential to produce DA whenever abiotic or biotic conditions trigger it (*Bates et al., 2018*; see section 1.3.1). It is thus important to precisely assess *Pseudo-nitzschia* biogeography at species level, especially in the context of global climate change. In fact, ocean warming is thought to directly trigger the proliferation of HABs, and on the other side it allows temperate toxic species to move poleward and colonize the more accessible Arctic and Antarctic ecosystems (*Lefebvre et al., 2016*). This invasion would in turn affect the trophic web hosted in these extreme environments that are changing faster than any other region on Earth and play a vital role in regulating the world's climate. Although a high resolution global-scale characterization of *Pseudo-nitzschia* spp. biogeography is needed, several studies have focused on regional and temporal variation within and among *Pseudo-nitzschia* species. This variability has been assessed both regarding the toxin production genes (*Bates et al., 2018*, and references therein) and at metabarcoding level. For instance, a recent study on temporal changes in *Pseudo-nitzschia* species living in sympatry in the Gulf of Naples (*Ruggiero et al., 2022*) found a high level of both inter- and intra-specific variability when looking at the hypervariable V4 region of the 18S rDNA marker of 33 species over 48 sampling dates covering a period of 3 years.

Species-specificity seems also to be the nature of the association between bacterial communities and *Pseudo-nitzschia* spp. Diatoms and bacteria are involved in complex interactions shaped by the long evolutionary co-existence in marine environment; a complete understanding of the ways and processes involved in these relationships is still far to be disentangled, but some advances have been made regarding diatom-bacteria and *Pseudo-nitzschia*-bacteria interactions in particular (reviewed in *Bates et al., 2018*).

For example, it is now clear that bacterial communities associated with *Pseudo-nitzschia* are highly host-specific (*Guannel et al., 2011*), and that in general toxic species host a lower bacterial diversity than non-toxigenic species (*Sison-Mangus et al., 2014, 2016*). The exchange of metabolites and chemical cues at the level of phycosphere, i.e., the microenvironment surrounding diatom cells that represents the planktonic equivalent of the rhizosphere in plants, is thought to play a significant role in mediating these interactions, which comprise mutualism, commensalism, competition, antagonism and parasitism (*Amin et al., 2012, 2015*; *Seymour et al., 2017*; *Zhou et al., 2018*). These species-specific complex interactions could take place both on diatoms' cellular surface and in the surrounding phycosphere, with bacteria involved appearing either as "algae-attached" or "free-living". Although crucial for many food webs, as well as holding a high potential for bioremediation of polluted environments (e.g., *Kahla et al., 2021*), these relationships have been poorly studied compared with higher plant–microbe associations. For example, it is not clear whether the free-living and the algae-attached communities are differentially specialized organisms adapted to one or the other lifestyle or whether they belong to a pool of generalist taxa with high plasticity that can easily switch their lifestyles depending on the most favourable conditions. In particular, while some studies reported differences between communities of free-living and attached bacteria (*Mohit et al. 2014*; *Zhang et al. 2016*; *Trano, 2021*), a variable degree of overlap was also observed (*Hollibaugh et al. 2000*; *Ghiglione et al. 2007*; *Rieck et al. 2015*; *Liu et al. 2018*). One way to study species-species interactions in natural environments are correlation networks, that are useful tools to uncover positive and negative relationships between different species, also accounting for the importance of environmental factors (*Barberan et al., 2012*; *Eiler et al., 2012*; *Needham et al., 2013*; *Pearman et al., 2016*; *Mikhailov et al., 2019*; *Xie et al., 2021*). Although positive correlations do not mean necessarily mutualism and, similarly, negative correlations are not unequivocal sign of competition, network analysis can still help to the study of communities and the interactions between their components in their natural habitat (*Ma et al., 2020*, *Krug et al., 2020*; *Xie et al.,*

2021), shedding light on the abiotic and biotic controls that shape microbial communities at sea.

The present chapter aims at first depicting the biogeographic distribution of *Pseudo-nitzschia* species. Given the above-mentioned impact of reference datasets and the risks of underestimating diversity when using clustered or denoised reads at low taxonomic level explorations, I will use a taxonomically curated reference database built for inter-specific and intra-specific variation analysis within *Pseudo-nitzschia* genus (*Ruggiero et al., 2022*) and I will look directly at single ribotypes instead of processed OTUs or ASVs. I will explore *Pseudo-nitzschia* biogeography and ecology at global scale with an unprecedented taxonomic resolution; for this exercise I will use not only V4-based information provided by *Tara* Oceans, but also data from other samples included in the UniEuk project and I will integrate the biological information with environmental parameters extracted from the World Ocean Atlas (WOA18; *Boyer et al., 2018*) to explore *Pseudo-nitzschia* species auto-ecology. In the second part of the chapter, the new built data set will be used to build correlation networks to study *in silico* interactions between *Pseudo-nitzschia* spp. and bacteria sampled during *Tara* Oceans and *Tara* Oceans Polar Circle expeditions. I will analyse correlation networks built using bacteria found in the free-living as well as in the particle- attached size fraction and highlight differences and similarities between the two communities.

4.2 Materials and Methods

4.2.1 *Pseudo-nitzschia* biogeography

The global distribution of *Pseudo-nitzschia* spp. was assessed using metabarcoding data from a subset of the projects included in the community-based UniEuk project. *Tara* Oceans and *Tara* Oceans Polar Circle expeditions data were included in the analysis as they are part of UniEuk. A complete list of the projects used is

shown in Table 4.1. For the selected projects I analysed metabarcoding data of the hypervariable V4 region of the 18S rDNA.

To generate the distribution table, I used a high-quality V4 data reference containing a selection of taxonomic validated *Pseudo-nitzschia* sequences from *Ruggiero et al. (2022)*. Given the lack of annotation of the 18S sequences for many species, the reference database was not representative of the whole genus; in particular, it was composed of 37 *Pseudo-nitzschia* sequences for a total of 33 species. Three distinct references were included for the toxic *P. galaxiae*, one for each of the three morphotypes and genotypes observed for this species in the Gulf of Naples (*Cerino et al., 2005*; *Ruggiero et al., 2015*); two different sequences belonged to *P. multiseries* and two to *P. australis*. Moreover, the dataset contained five groups of species with identical V4 sequences (groups ID_1 – ID_5). Group ID_1 includes four different species, namely *P. arenysensis*, *P. dolorosa*, *P. arctica* and *P. subcurvata*; Group ID_2 includes five species, i.e., *P. americana*, *P. caciantha*, *P. circumpora*, *P. granii* and *P. linea*. The other three groups sharing an identical haplotype are represented by two species each: group ID_3 included *P. australis* and *P. multiseries*, group ID_4 included *P. brasiliana* and *P. lineola*, while group ID_5 included *P. kodamae* and *P. lundholmiae*. It is worth mentioning that both *P. australis* and *P. multiseries* are represented both by single sequences and by the sequences included in the group ID_3. I reported the complete list of sequences, summarising the information from *Ruggiero et al. (2022)* in Table 4.2.

The taxonomically curated V4 reference sequences presented in the above-described database were used as queries for a local BLAST against the UniEuk V4 dataset. UniEuk metabarcoding reads are normally denoised to build ASVs; however, for the following analysis, I directly used the metabarcode haplotypes, i.e., the reads after quality filtering but before undergoing the denoising process that leads to the formation of ASVs; in this way I could increase the resolution of the final data set (see paragraph 4.1.). These unpublished data were provided by Dr. Colomban de Vargas (CNRS France). In order to map the sequences at the species

level with the maximum reduction of ambiguity, I set a threshold of 100% identity and retained only the metabarcoding hits with a query coverage with the reference of at least 380 bp. I then generated a table with the occurrence of each species in each sample, filtering only samples of surface water (≤ 10 meters). The heterogeneity of the UniEuk dataset, that comprises distinct projects with different sampling strategies, did not allow me to retrieve the information of size classes for each project thus preventing me from using the abundance information. The obtained occurrence table was used to map the global-scale *Pseudo-nitzschia* species distribution. In particular, I first plotted the information of the richness of each sample of the obtained dataset, and then the occurrence of each species in each sample.

Table 4.1. List of projects included in UniEuk used for the global-scale Pseudo-nitzschia species biogeography and ecology assessment.

Project ID	Description
PRJEB6610; PRJEB9737	Amplicon sequencing of *Tara* Oceans and *Tara* Oceans Polar Circle DNA samples corresponding to size fractions for protists.
PRJNA311248	DNA and RNA based tag sequences raw sequence reads
PRJNA385736	Marine metagenomes (The Australian Microbiome Initiative)
PRJNA391338	SPOT April V4 tag sequences Raw sequence reads
PRJEB24314	Diversity of marine microbial eukaryotes in Nares Strait in summer 2014 and 2016
PRJEB25353	Diversity of picoeukaryotes in Northern California coastal water
PRJEB25232	Samples of eukaryotic plankton sequenced at the V4 rDNA region
PRJEB23873	Environmental MicroEukaryote sampling in Nago Bay, Okinawa

PRJEB23005	Microbial diversity of the central Arctic Ocean
PRJNA391477	Seasonal HOT cruise vertical profile 18S tag (2014-2015) Raw sequence reads

Table 4.2. List of species and GenBank accession ID of the sequences of the reference dataset used. Modified from Ruggiero et al. (2022).

Species	GenBank accession	Reference
P. allochrona	KJ608076	Percopo et al. (2021)
P. americana	MG799146	Unpublished
P. arenysensis	OK135731	Ruggiero et al. (2022)
P. arctica	JF794046	Percopo et al. (2016)
P. australis	AM235384	Koester et al. (2013)
P. australis	GU373961	Fitzpatrick et al. (2010)
P. batesiana	KP708989	Lim et al. (2016)
P. brasiliana	KP708990	Lim et al. (2016)
P. caciantha	KP708992	Lim et al. (2016)
P. calliantha	OK103848	Ruggiero et al. (2022)
P. hasleana	JN091716	Lundholm et al. (2012)
P. circumpora	KP708994	Lim et al. (2016)
P. cuspidata	KP708995	Lim et al. (2016)
P. decipiens	KP708996	Lim et al. (2016)
P. delicatissima	KJ608075	Percopo et al. (subm)
P. dolorosa	OK103849	Ruggiero et al. (2022)
P. fraudulenta	KJ608077	Percopo et al. (subm)
P. fukuyoi	KP708997	Lim et al. (2016)
P. galaxiae	OK103850	Ruggiero et al. (2022)
P. galaxiae	KJ608078	Percopo et al. (2021)

P. galaxiae	KJ608079	Percopo et al. (2021)
P. granii	JN934671	Percopo et al. (2016)
P. heimii	JN091727	Unpublished
P. kodamae	KP709000	Lim et al. (2016)
P. linea	OK103851	Ruggiero et al. (2022)
P. lineola	JN091717	Unpublished
P. lundholmiae	KP709001	Lim et al. (2016)
P. mannii	KJ608080	Percopo et al. (2021)
P. micropora	KP709003	Lim et al. (2016)
P. multiseries	AM235380	Unpublished
P. multiseries	U18241	Manhart et al. (1995)
P. multistriata	OK103852	Ruggiero et al. (2022)
P. pseudodelicatissima	KJ608082	Percopo et al. (2021)
P. pungens	KP709004	Lim et al. (2016)
P. sabit	KP709005	Lim et al. (2016)
P. subcurvata	KX253952	Moreno et al. (2018)
P. turgidula	FJ222752	Unpublished

4.2.2 *Pseudo-nitzschia* autoecology

The occurrence table generated using the curated V4 reference and the UniEuk dataset was used to link the presence and absence of species to environmental parameters extracted from the World Ocean Atlas database (WOA18; *Boyer et al., 2018*). These climatological data are relative to the average monthly values. In particular, I plotted histograms of the distribution of two physical parameters, i.e., temperature and salinity, and the three main macronutrients for diatoms, i.e., nitrate, phosphate and silicate, across the UniEuk samples. To do that, values of nutrients were considered only if above a selected threshold: in particular, NO_3^- and $Si(OH)_4$ concentrations lower than 0.1 µmol/L where set as equal to 0.1; for

PO_4^{3-} a threshold of 0.05 was used, with lower values set as equal to 0.05. Nutrients concentrations were plotted in logarithmic scale. I then constructed boxplots to evaluate the optimal range of values of environmental parameters for each *Pseudo-nitzschia* species. In order to describe the community structure using the environmental parameters, I performed a Canonical Correspondence Analysis (CCA), that uses the analysis of variance in the community matrix to find the best set of environmental factors that explain the observed structure of community. I used the *adonis* function, from the R package *vegan* (*Oksanen et al., 2013*), that fits linear models to distance matrices and uses a permutation test with pseudo-F-ratios. Given the use of occurrence information, the dissimilarity matrix was based on the Jaccard distance, appropriate for presence-absence data sets. Only the environmental variables that significantly explained the community matrix, i.e., showing an associated p-value equal or below a cut-off of 0.01, were drawn in the CCA plot. The same procedure described in section 4.2.1 was used to build a table of *Pseudo-nitzschia* species abundance across sampling stations of *Tara* Oceans and *Tara* Oceans Polar Circle expeditions. The homogeneity of *Tara* samples in terms of sampling strategy allowed me to retrieve not only the occurrence of species but also the abundance information of each *Pseudo-nitzschia* species at each sampling station for different depths and size fractions. From this data set, sequences with less than 3 reads (i.e., detected less than three times in the whole dataset) and appearing in less than 2 samples were discarded. Moreover, only samples with at least 10 *Pseudo-nitzschia* reads in total were retained for the following analysis. *Tara* Oceans and *Tara* Oceans Polar Circle data are organized in size fractions; the microplanktonic (20-180 µm) is by far the most used size class when analysing diatoms, but the importance of nanoplanktonic (identified as 3-20/5.20µ) diatoms has been recently advocated (*LeBlanc et al., 2018*). *Pseudo-nitzschia* is a highly heterogeneous genus, with elongated, fusiform cells, whose size ranges from 2 to 8 µm in width and from 40 to 175 µm in length. Therefore, although it is possible that some cells perpendicularly pass through filters and are retrieved in the nanoplanktonic class, the selected size class for analysing *Pseudo-*

nitzschia species in the present chapter has been the microplanktonic fraction, i.e. the 20-180µ. This dataset was used to explore *Pseudo-nitzschia* abundance and distribution in the surface global ocean.

4.2.3 *Pseudo-nitzschia* synecology

In order to investigate the co-variation patterns of *Pseudo-nitzschia* species with bacteria, two different data sets were merged. First, the data set obtained as described in section 4.2.2., i.e., the one representing *Pseudo-nitzschia* species abundance and distribution in the microplanktonic size fraction in the surface *Tara* Oceans and *Tara* Oceans Polar Circle sampling stations. Second, the prokaryotic metabarcoding data that targeted the V4 and V5 region of the 16S rDNA; in particular, I used the rarefied tables at 20,000 reads/sample that were built to correct for sequencing depth. Diatom-bacteria interactions are known to occur both with bacteria attached to the diatom surface as well as with bacteria appearing in the water surrounding algae (see section 4.1.). In order to discriminate between putatively algae-attached bacteria and free-living ones, I separately considered for the subsequent analysis two different prokaryotic size classes: prokaryotic OTUs occurring in the same size fraction used for *Pseudo-nitzschia* species, i.e., 20-180 µm, were considered as particle-attached bacteria, while bacteria found in the 0.22-3 µm size class were defined as free-living.

4.2.3.1 Correlation networks

The first step to construct the correlation-based network has been the calculation of the pairwise correlation coefficients between every pair of vectors of the considered data sets. In particular, I calculated the Spearman correlation coefficients between the abundance vectors of each *Pseudo-nitzschia* species and each bacterial OTUs across *Tara* samples. I converted the absolute number of prokaryotic and *Pseudo-nitzschia* read counts into relative abundance by dividing by the total counts of bacteria or *Pseudo-nitzschia* species in each sample,

respectively (Total Sum Scaling, TSS). Since the scope of the network is to visualize patterns of covariation of species across several samples, I considered here only *Pseudo-nitzschia* species and bacterial OTUs occurring in at least five stations. This procedure, although important for statistical robustness, inevitably reduces the total number of stations targeted.

The correlation matrices obtained, representing the *Pseudo-nitzschia* – free living and the *Pseudo-nitzschia* – attached bacteria covariation, were then filtered in order to retain only significant positive strong correlations. In particular, the p-value threshold, that reflects the significance of the correlation, was set equal to 0.05. Moreover, I selected only correlation coefficients (r) with a magnitude greater than 0.5 and positive ($r > 0.5$). The filtered correlation matrices were then plotted using the R package *igraph* (*Csardi and Nepusz, 2006*); the resulting networks illustrate the significant positive co-variations in abundance between each *Pseudo-nitzschia* species and each prokaryotic OTU.

4.2.3.2 Number of interactions and species-specificity

The degree, defined as the number of edges of each node, was calculated for each *Pseudo-nitzschia* species of both networks. This value, that reflects the number of putative interactions between each diatom and each prokaryotic OTU, was evaluated in the context of the mean relative abundance of each *Pseudo-nitzschia* species ID in each sampling station and compared to each diatom's occupancy, i.e., the number of distinct stations in which each species ID is found. In order to test whether *Pseudo-nitzschia* species are preferably associated with a unique set of bacteria or with prokaryotic OTUs that are also robustly co-varying with other species, I firstly calculated a species-specificity index. The index was obtained by dividing the number of bacteria uniquely linked to one *Pseudo-nitzschia* species by the total degree of the same species. Furthermore, I explored the taxonomic composition of the prokaryotic community linked to each *Pseudo-nitzschia* species in both networks.

4.2.3.3 Free-living and particle-attached Bacteria

With the aim of exploring whether there was a difference between the prokaryotic communities in the free-living (FL) or in the particle-attached (PA) size fractions, I calculated the relative proportion of prokaryotic OTUs associated with each *Pseudo-nitzschia* species that were uniquely found in the FL-based or in the PA-based network, together with the assessment of the bacteria that were common to both networks. Three categories were obtained: bacteria that are uniquely found in the FL-based network, bacteria exclusive to the PA-based network and bacteria that are positively co-varying with *Pseudo-nitzschia* species in both networks. The three categories of bacteria were explored from a taxonomical point of view, by assessing the similarities and differences at different taxonomical ranks, i.e., phylum, family and genus. Venn diagrams were built using the R package *venn* (*Dusa, 2020*) to evaluate the overlap between groups; moreover, bar plots were used to investigate the most abundant phyla, families and genera for the three categories of bacteria.

4.3 Results and Discussion

4.3.1 *Pseudo-nitzschia* biogeography

The final UniEuk data set analysed included 1055 samples belonging to a total of 10 projects (see Table 4.1.). The 25 "species IDs" corresponded to the 33 *Pseudo-nitzschia* species represented by the reference dataset; this mismatch is due to the fact that five IDs identify more than one species each. The histogram in Fig. 4.1. shows the number of samples in which each ribotype is found. The most widespread were, as expected, the ones that represent several species, like the cases of the groups ID2 (*P. americana*, *P. caciantha*, *P. circumpora*, *P. granii* and *P. linea*) and ID1 (*P. arenysensis*, *P. dolorosa*, *P. arctica* and *P. subcurvata*). The least widespread species were *P. mannii*, *P. decipiens*, *P. allochrona*, *P. calliantha*, and the two ribotypes that identify *P. multiseries* and *P. australis*, two species that show high intraspecific variability of the V4 rDNA gene, being also included in the group ID_3. *P. galaxiae* is represented by three ribotypes, out of which two occurred in a high number of samples and one had a narrow distribution.

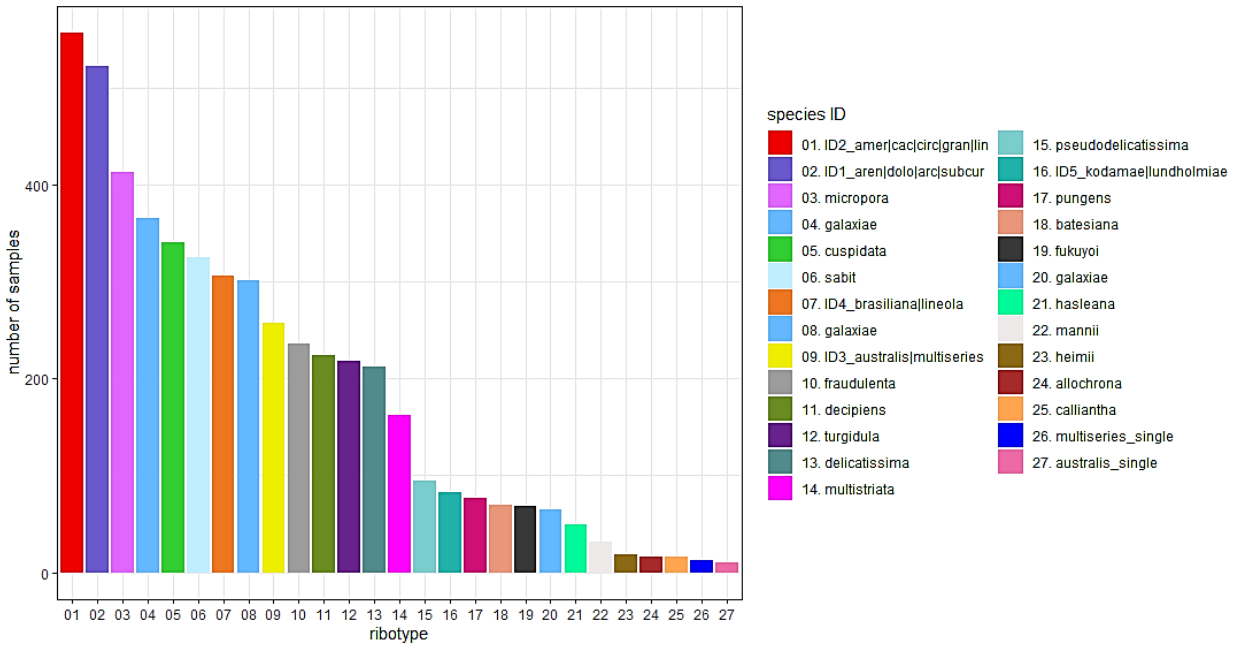

Figure 4.1. Histogram showing the number of UniEuk samples in which each ribotype is found. Colours represent the species ID.

Patterns of richness, expressed as the number of distinct species ID in each sample, are shown in Fig. 4.2. Richness ranged from 1 to 20, with the majority of samples showing a low value. Only a few samples displayed a high biodiversity (> 16 species): one was located at the Gibraltar Strait, one in the Southern Indian Ocean, and one in the North Pacific Ocean, in a region characterized by the Californian current. However, the coordinates of the latter represented the one of three samples (overlapped in the Figure) that were the only three belonging to the project PRJEB25353, a collection of picoeukaryotes. Picoeukaryotes are defined as holding a maximum size of 3 µm, while *Pseudo-nitzschia* species, and diatoms in general, are much larger; it is therefore interesting that the Californian samples held a high diversity (12-15 species), that however most probably represented fragments of *Pseudo-nitzschia* broken cells that remained suspended in the water column after cell death.

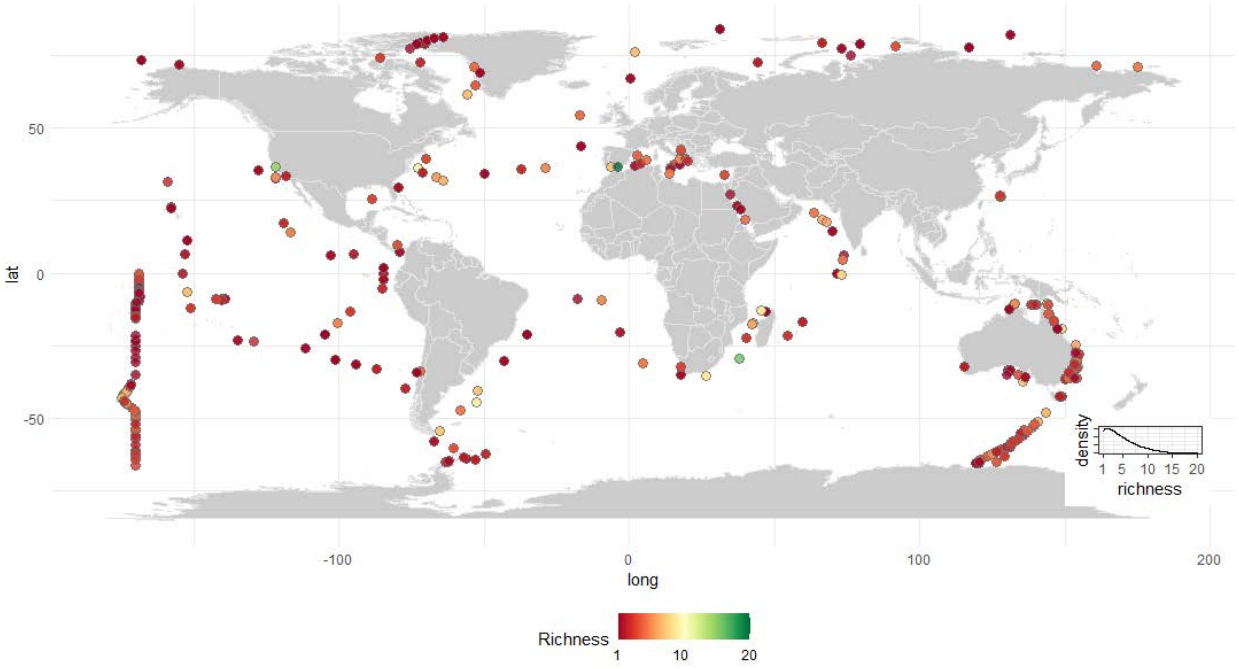

Figure 4.2. Spatial distribution of UniEuk samples with colours indicating the total richness, i.e., the number of Pseudo-nitzschia species ID. The density plot (bottom right) indicates the distribution of richness across samples.

The most recent assessment the global distribution of *Pseudo-nitzschia* spp. is found in *Bates et al. (2018)*. However, this comprehensive review did not include the recently described *P. allochrona* (*Percopo et al., 2021*), one of the cryptic species belonging to the *P. delicatissima*-species complex. This species was so far documented only in the Mediterranean Sea (*Percopo et al., 2021*), being present in the Gulf of Naples (*McDonald et al., 2007*; *Lamari et al., 2013*; *Ruggiero et al., 2015*), the Ionian (*Percopo et al., 2021*) and the Adriatic Sea (*Pugliese et al., 2017*; *Arapov et al., 2020*; *Giulietti et al., 2021*). The present results (Fig. 4.3.) confirm the occurrence of *P. allochrona* in the Mediterranean Sea; interestingly, the global scale extension provided by UniEuk allowed the detection of this species also in the Western Mediterranean Sea and at the Gibraltar Strait, and in a sample just north of the Panama Canal. Besides the novelty of *P. allochrona*, general trends and exceptions emerged from the global-scale analysis of *Pseudo-nitzschia* species.

P. australis and *P. multiseries* were a peculiar case. Among the most toxic *Pseudo-nitzschia* species, their biogeography has important implications for ecosystem health at global scale. Both species show high intraspecific variation when considering the V4 region of the 18S rDNA such that each of them was present in the reference database with two different ribotypes. However, one of the two ribotypes of each species was identical between the two. This means that three different ribotypes described *P. australis* and *P. multiseries*: one identified *P. australis* only (named *P. australis*_single), one represented *P. multiseries* only (*P. multiseries*_single) and one did not discriminate between the two (ID_3_*australis*|*multiseries*). *P. australis*_single had a clear cold-water pattern: it occurred in three samples in the California Current, a cold-water Pacific Ocean current that flows southward along the western coast of North America representing one of the five major coastal currents affiliated with strong upwelling zones. This species also occurred in a sample on the south eastern coast of Australia and in the Arctic Ocean. *P. multiseries*_single occurred in the Arctic Ocean (Labrador Sea), in the Californian upwelling region, in one sample in the Southern

Atlantic Ocean, but also in two tropical Pacific Ocean samples. ID_3_*australis*|*multiseries* was widespread, thus probably hiding unknown level of cryptic diversity. Besides the case of *P. australis*, species showing a specific cold-water distribution did not emerge from the global biogeography investigation: species that occurred in polar areas were generally found also in tropical and temperate waters. However, it must be noticed that I could not identify the two polar species *P. arctica* and *P. subcurvata*. The lack of resolution power of the hypervariable rDNA region V4 in distinguishing among *P. arenysensis*, *P. dolorosa*, *P. arctica* and *P. subcurvata* (pulled together in group ID_1) limited the accuracy of the biogeography assessment. Interestingly, these four species are all non-toxic, but occupy different regions of the ocean. While *P. arenysensis* and *P. dolorosa* share the occurrence in Australian Malaysian waters and in Mediterranean Sea, *P. arenysensis* was also found in California, *P. dolorosa* occurs in Argentina, Mexico and Namibia, while *P. arctica* and *P. subcurvata* are apparently two truly polar species (see *Bates et al., 2018*, and references therein). *P. arctica*, a recently discovered species, is in fact found in Arctic Ocean, while *P. subcurvata* has been solely reported for the Southern Ocean. It is interesting that such different species show the identical V4. The DNA sequence identity in the V4 sequence of an Arctic and a Southern Ocean species could erroneously lead researchers to suggest the existence of bipolar OTUs if these two species were clustered together using the V4 marker even at high similarity thresholds.

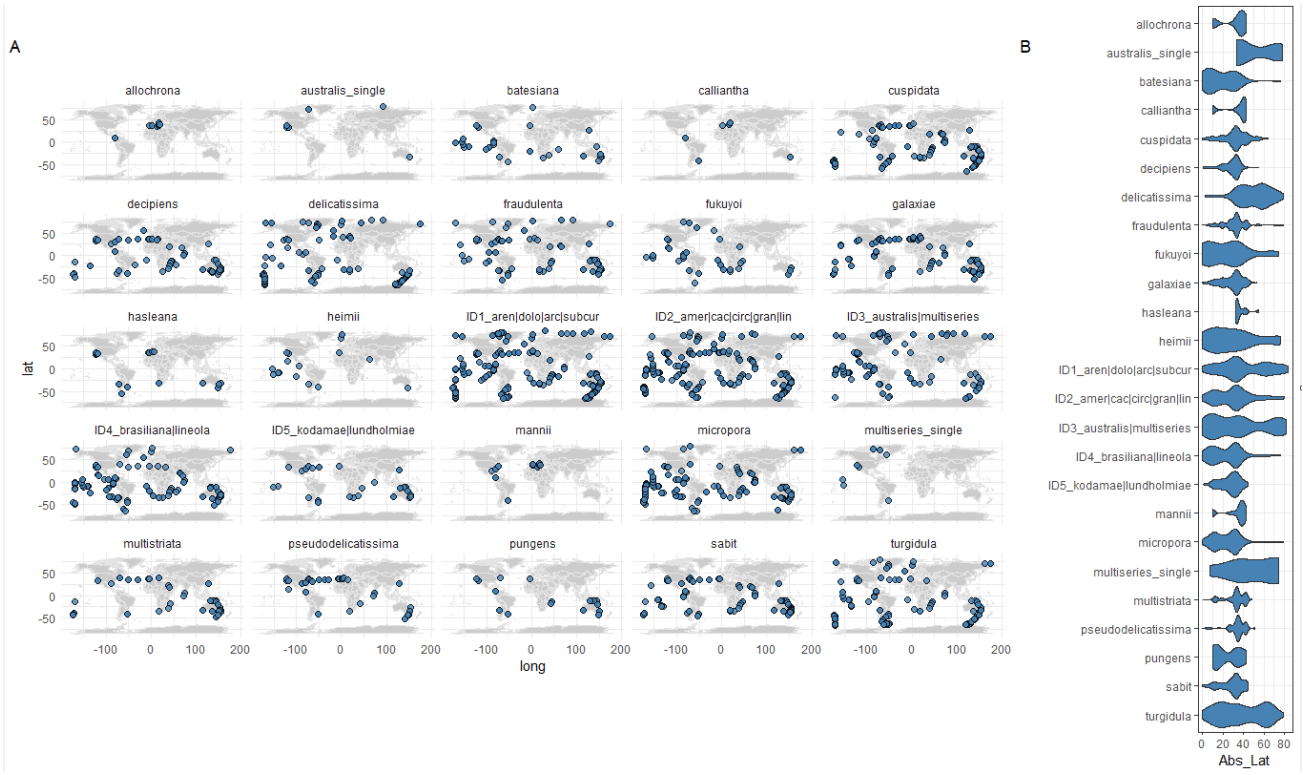

Figure 4.3. Overview on Pseudo-nitzschia species biogeography using presence-absence data. A) Geographical distribution of each species. B) Distribution of species occurrences along the latitudinal gradient expressed in absolute values.

If, on the one hand, polar-endemic species did not emerge from the global biogeography analysis, on the other hand some species showed a clear non-polar pattern. For example, several species had a temperate-tropical distribution. In particular, *P. allochrona*, *P. calliantha*, *P. mannii* and *P. pungens* displayed a tropical-subtropical distribution. The extent of their latitudinal ranges went from a minimum of ±10° to a maximum of ±40°, thus showing a narrower range compared to the majority of *Pseudo-nitzschia* species. However, while *P. allochrona* was only found above the equator, the other three species also occurred in the austral hemisphere. The group ID5, that pulled together *P. lundholmiae* and *P. kodamae*, showed a pattern similar to the above-mentioned species, with a range of distribution that extended from -44° to 36° of latitude. The species with the narrowest range of absolute latitudinal distribution was *P. hasleana*, that showed a striking symmetry respect to the Equator. Other species occurred across a wider

geographical range, with latitude that extended from tropical to temperate values. This is the case of *P. decipiens*, *P. galaxiae*, *P. multistriata*, *P. pseudo-delicatissima* and *P. sabit*. *P. cuspidata* was also mainly tropical and temperate, but it occurred in some samples in the Southern Ocean (-64° of latitude), being thus also capable to survive in colder waters. However, this happened only in austral hemisphere, with no occurrence of this species in the Arctic Ocean. *P. batesiana* and *P. heimii* shared a similar distribution, reaching cold waters especially in the northern hemisphere, never approaching the inner part of the Arctic Ocean but occurring in the Western Eurasian Basin, i.e. the Atlantic-influenced region of the Arctic, also characterized by higher concentrations of nitrates and phosphates and a lower level of iron. Cosmopolitan species that thrive at every latitude and thus in different oceanographic conditions included *P. delicatissima*, *P. fraudulenta*, *P. fukuyoi*, *P. micropora* and *P. turgidula*, together with the already described groups of ID1 and ID3, and the ID2 group, that includes *P. americana*, *P. cacianthа*, *P. circumpora*, *P. granii* and *P. linea*.

4.3.2 *Pseudo-nitzschia* autoecology

Pseudo-nitzschia spp. autoecology has been investigated through the visualization of each species' tolerance ranges for two physical parameters, i.e., temperature and salinity, and three macronutrients, i.e., nitrate, phosphate and silicate, as extracted from the World Ocean Atlas (WOA18; *Boyer et al., 2018*). The same parameters were used to build a Canonical Correspondence Analysis (CCA), that allowed to explain the community structure using environmental factors. This is a complementary step that allows to characterize the sample points not only through their latitudes and longitudes, but also looking at the environmental factors that shape the oceanographic conditions that the species were encountering at that time in that location. As seen from the biogeography assessment, the majority of species displayed similar ranges of temperatures. Cold waters were tolerated by the cosmopolitan *P. turgidula* and *P. delicatissima*, that also showed the preferences

for high nitrate and phosphate, two nutrients whose concentrations are higher in the Southern Ocean. Moreover, *P. australis* and *P. multiseries* showed occurrence in cold waters both when looking at the unique ribotypes and at the group level (*P. australis*_single, *P. multiseries*_single and ID3_*australis*|*multiseries*, respectively). As shown in Fig. 4.3., the cold-adapted species were not unique to cold environments, but rather appeared as eurythermal species, i.e., capable of tolerating a wide range of temperatures. However, the temperature histogram in Fig. 4.4 shows a dip at 10 ˚C that suggests an under sampling of temperate waters that could have affected results.

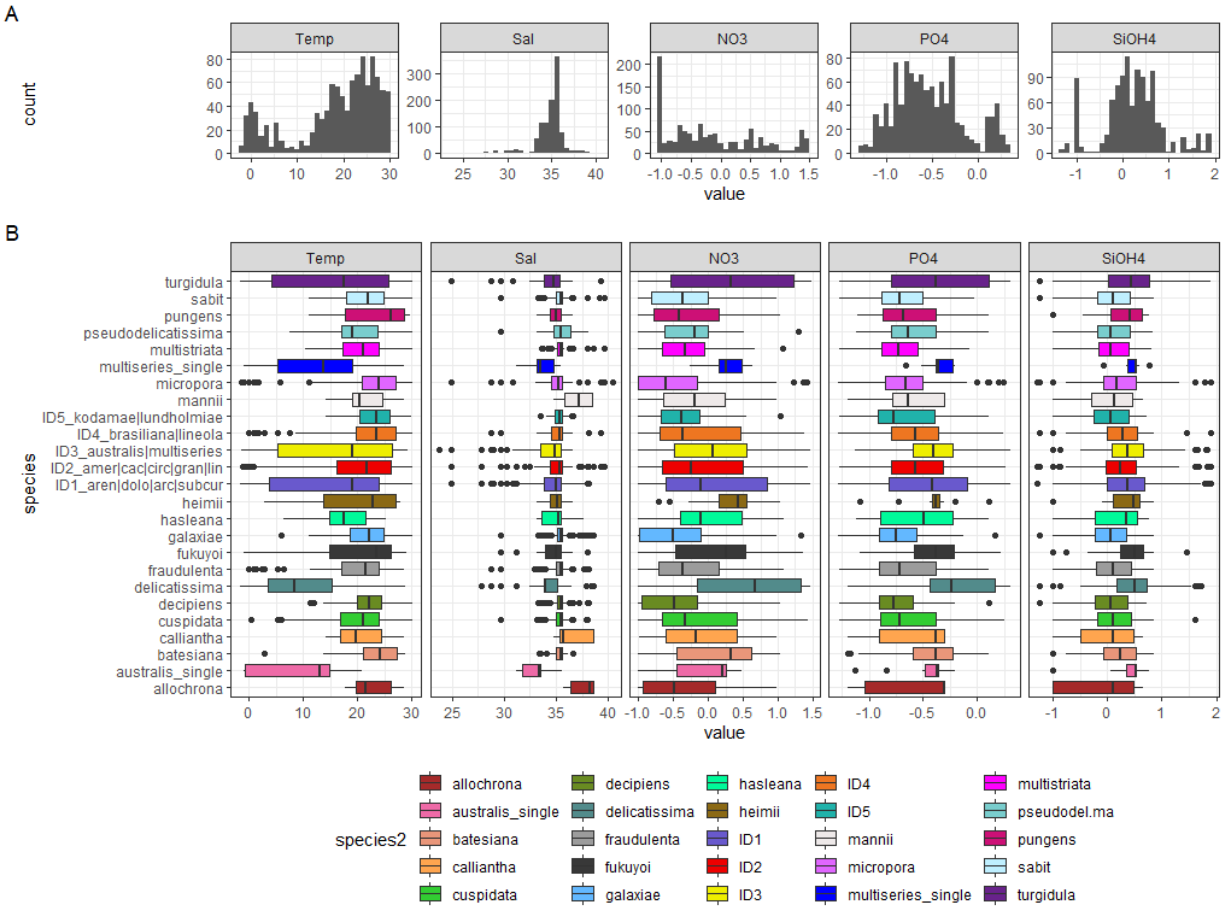

Figure 4.4. A) Histograms of the distribution of values of environmental parameters across samples. B) Boxplots showing the range of values tolerated by each Pseudo-nitzschia species ID across UniEuk samples. In both A) and B) plots, values of nutrients below selected thresholds were set as equal to them (see par. 4.2.2.) and then plotted in logarithmic scale.

All the five environmental descriptors used in the CCA were found to be significant (p-value ≤ 0.01) and were thus displayed in the ordination plot in Fig. 4.5. The CCA helped to disentangle the ecological nature of the biogeographical patterns observed and helped the distinction of a group of species occurring at low temperatures (*P. australis*_single and *P. delicatissima*) and high nutrient values (*P. multiseries*_single, *P. turgidula*, ID1). An effect of river outflows could be one reason for the separation of species at low salinity levels, together with the low salinity of polar waters. The arrows corresponding to the nitrate, phosphate and silicate were all very close to each other, thus suggesting a high correlation and preventing to hypothesize the role of each nutrient in shaping *Pseudo-nitzschia* communities.

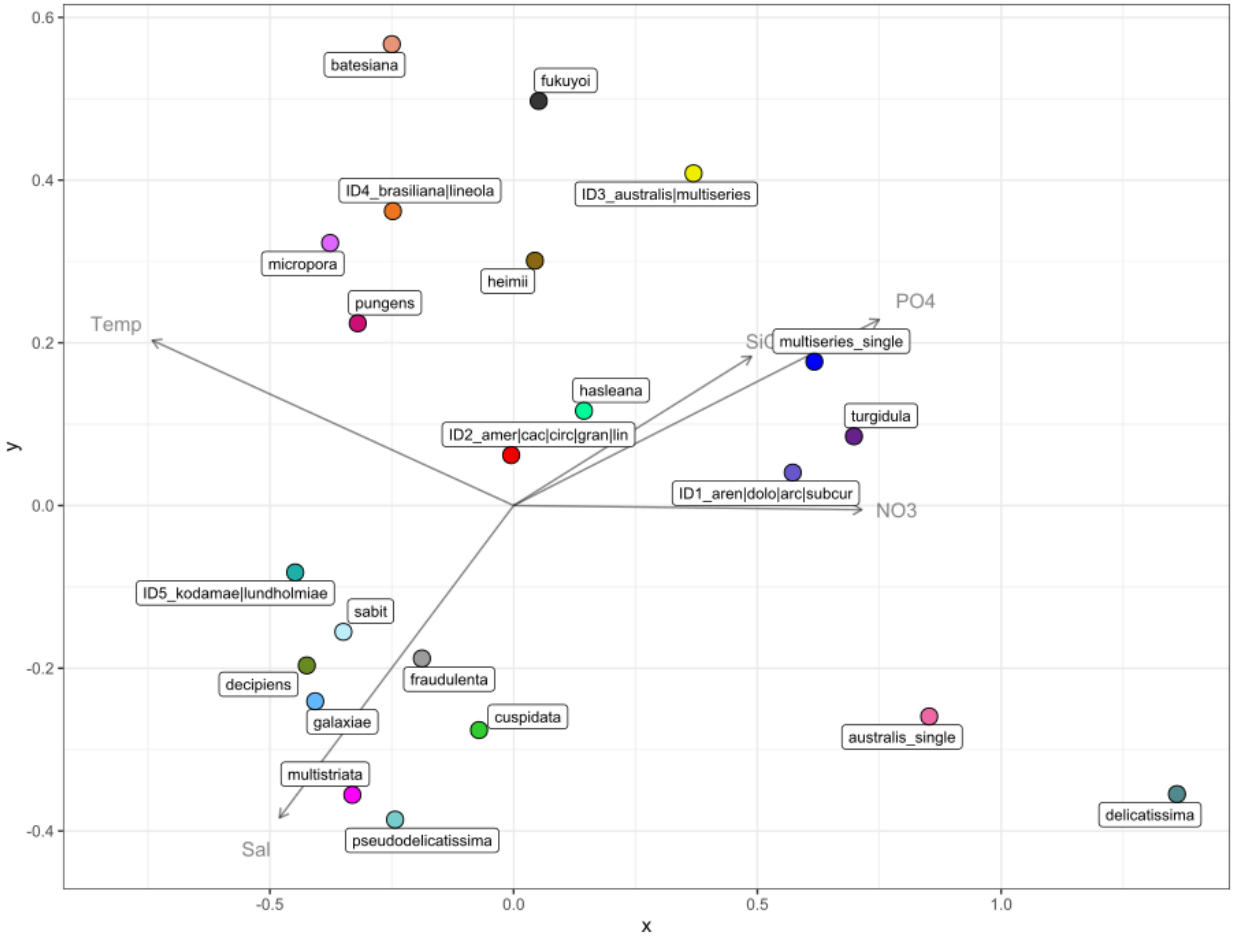

Figure 4.5. CCA ordination plot depicting the relationship between environmental parameters and Pseudo-nitzschia species community structure at global scale. Three outlier species were removed, i.e., P. allochrona, P. calliantha, P. mannii.

Pseudo-nitzschia biogeography was also studied using the information on species abundance across *Tara* Oceans and *Tara* Oceans Polar Circle sampling stations. *Pseudo-nitzschia* species were extracted from surface samples corresponding to the 20-180µm size fraction (microplankton). A visualization of the biogeography of species using the *Tara* surface microplanktonic data is shown in Fig. 4.6. The distribution generally matched the one obtained with UniEuk dataset, but here the assessment of the relative contribution of each "species ID" at each site was made possible by the integration of the relative abundance information. The ID2 (*P. americana*|*caciantha*|*circumpora*|*granii*) and ID4 (*P. brasiliana*|*lineola*) groups dominated, together with the cosmopolitan *P. micropora*, in large part of the Pacific Ocean, as well as the Northern Indian Ocean and the Eastern South Atlantic samples, with ID2 being also the predominant species group in one station at the Eastern Mediterranean Sea (station 30), in a Southern Atlantic Ocean sample (station 68) and at the transition region between the North Atlantic and the Arctic Ocean (station 158).

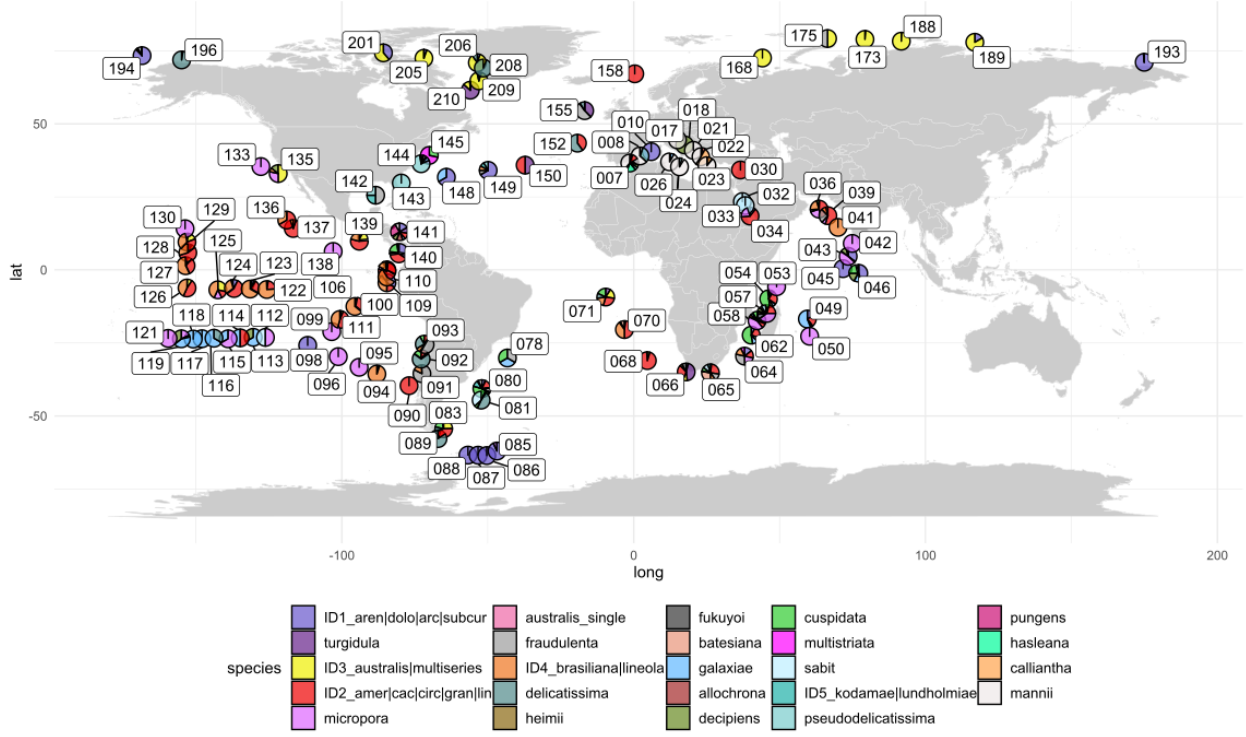

Figure 4.6. Abundance and distribution of Pseudo-nitzschia spp. across Tara Oceans and Tara Oceans Polar Circle surface stations for the microplanktonic size fraction (20-180µ).

The latter was dominated by the group ID3, constituted by *P. australis* and *P. multiseries*. *P. fraudulenta* was also abundant in this region, and three stations in the Pacific-influenced Arctic waters (stations 193-196) displayed a different community composition, with the dominance of the group ID1 (that includes *P. arctica*) and *P. delicatissima*. The latter was highly abundant also in a station in Labrador Sea (station 208) as well as in some subpolar samples in the austral hemisphere (stations 81 and 89). The group ID1, including tropical-temperate species like *P. arenysensis* and *P. dolorosa*, showed high abundance in one Mediterranean station (10) as well as in a South Atlantic and two North Atlantic samples (98, 148 and 149, respectively). It was also almost the only species to contribute to the total *Pseudo-nitzschia* spp. abundance in the Southern Ocean samples (stations 85-88); however, its signal in this station was probably provided by *P. subcurvata*. The Arctic Ocean also hosted *P. turgidula*, that dominated the southern sample of the

Labrador Sea (station 210); this species constituted the 50% of the total *Pseudo-nitzschia* abundance a North Atlantic sample (station 150) and represented a high fraction in a location sampled along the North Atlantic Drift (station 155). Interestingly, the *P. australis*|*P.multiseries* group, together with *P. delicatissima*, also appeared in a subpolar sample in the South Atlantic Ocean (station 083), and it constituted half of the total *Pseudo-nitzschia* abundance in the cold waters that inhabit the Californian upwelling region (station 135). The ribotype that groups together these two toxic species occurred at all latitudes (Fig. 4.3) but it seemed to dominate cold and polar waters, whose monitoring for the presence of toxic species has been strongly suggested (*Lefebvre et al., 2016*). Other regions were more prone to host a high diversity of species, as shown for several Indian Ocean samples. The Mediterranean Sea was dominated by *P. mannii*, while *P. multistriata* was the dominant species in station 145.

4.3.3 *Pseudo-nitzschia* synecology

The above-described data set was used to build correlation networks with both Bacteria occurring in the 0.22-3 and 0.22-5 µm size class (free-living, FL) and with the ones appearing in the same size class as the diatom species, i.e., 20-180 µm. Although it is not possible to ascertain the real nature of the prokaryotic OTUs found in the large size fraction, I will refer to them as particle-attached (PA), to compare them with the FL bacteria.

The network in Fig. 4.7 is the graphical representation of the correlation between the relative abundances of free-living bacteria and *Pseudo-nitzschia* species occurring in the 20-180 µm size class. Choosing only strong positive correlations ($r > 0.5$) between diatoms and bacteria occurring in at least 5 stations led to a final selection of 17 *Pseudo-nitzschia* IDs and 2,268 prokaryotic OTUs, whose relative abundances have been correlated across 35 sampling stations covering the Southern Indian Ocean, the Pacific, the Atlantic and the Arctic Ocean (Fig. 4.7). The network shows a complex structure with many species co-variating with a specific

set of bacterial OTUs that were never found linked with other species. Remarkable examples of this behaviour are displayed by *P. australis|multiseries*, *P. delicatissima*, *P. turgidula* and *P. fukuyoi*, linked with a high number of prokaryotic nodes out of which a few were shared with other *Pseudo-nitzschia* species. The majority of the other species showed a more compact structure, being linked by the same set of correlated bacteria.

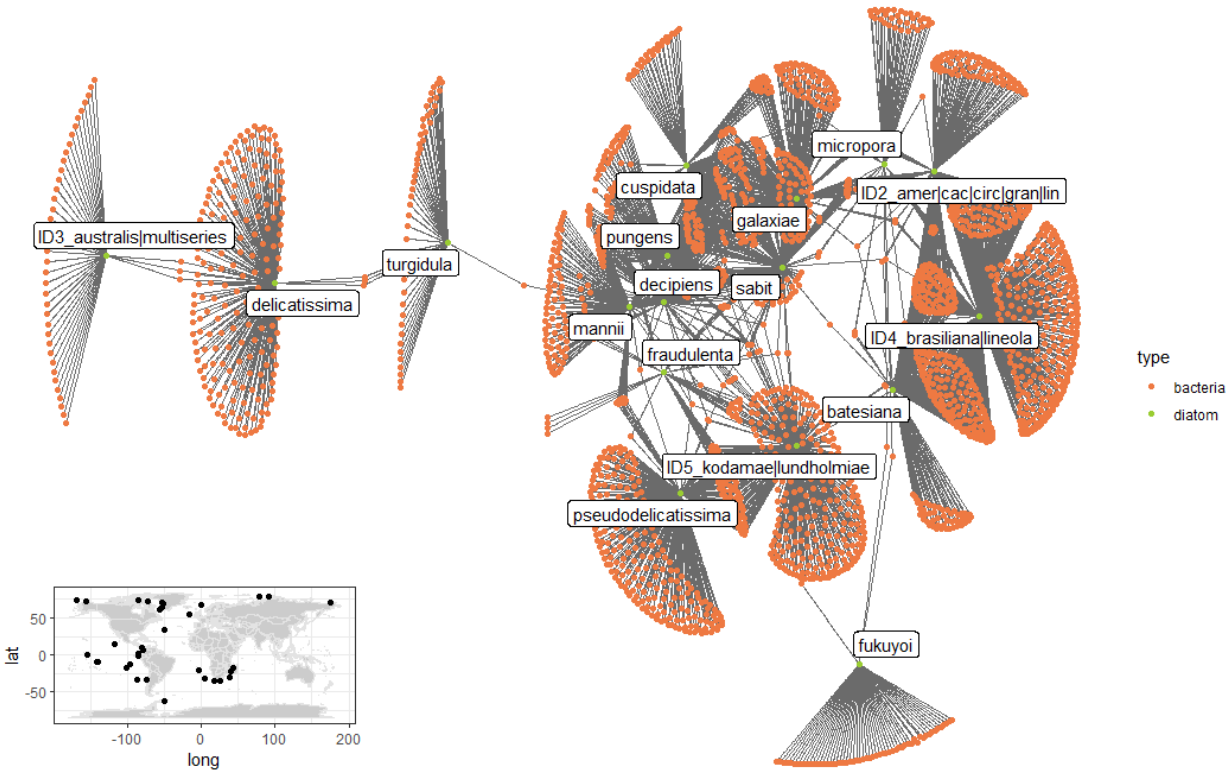

Figure 4.7. Correlation network visualizing significantly strong positive pairwise Spearman correlations between Pseudo-nitzschia species (green labelled dots) and free-living prokaryotic OTUs (orange dots) across 35 sampling stations (bottom-left). A total of 17 Pseudo-nitzschia species ID (20-180 µm) and a 2268 prokaryotic OTUs (0.22-3 µm) were used to build the network.

A high level of species-specific association also emerged from the network of putatively attached bacteria (Fig. 4.8), i.e., the one built from pairwise correlations of both prokaryotes and diatoms occurring in the 20-180 µm. The network consisted of 19 *Pseudo-nitzschia* "species IDs" and 2355 prokaryotic OTUs, whose abundances have been correlated across 40 sampling stations that covered a wide

geographical extension but, contrary to the previous network, were almost never found in the Arctic Ocean. Clouds of bacterial OTUs that were linked with one or two *Pseudo-nitzschia* species dominated the structure of this network, that seems overall composed by three main subgroups, each one linked to the others through only a few numbers of bacterial OTUs. One group was made by four species IDs that shared links with similar prokaryotic OTUs and showed a compact substructure in the global network; this group was composed by the species groups ID2 (*P. americana*|*caciantha*|*circumpora*|*granii*|*linea*), ID3 (*P. australis*|*multiseries*) and ID4 (*P. brasiliana*|*lineola*) together with *P. batesiana*. However, this result is probably driven by the presence of several distinct species that are forcedly grouped together. The second group consisted of *P. turgidula*, *P. cuspidata*, *P. sabit*, *P. decipiens*, *P. galaxiae*, *P. pseudodelicatissima*, *P. micropora*, *P. pungens*, *P. fukuyoi* and the species group ID5 (*P. kodamae*|*lundholmiae*). A single edge linked this group with the species group ID1 (*P. arenysensis*|*dolorosa*|*arctica*|*subcurvata*), that in turn was linked with other three bacteria. A third group, less compact, consisted of *P. fraudulenta*, *P. delicatissima*, *P. calliantha* and *P. mannii*. The latter two species shared a high number of bacteria.

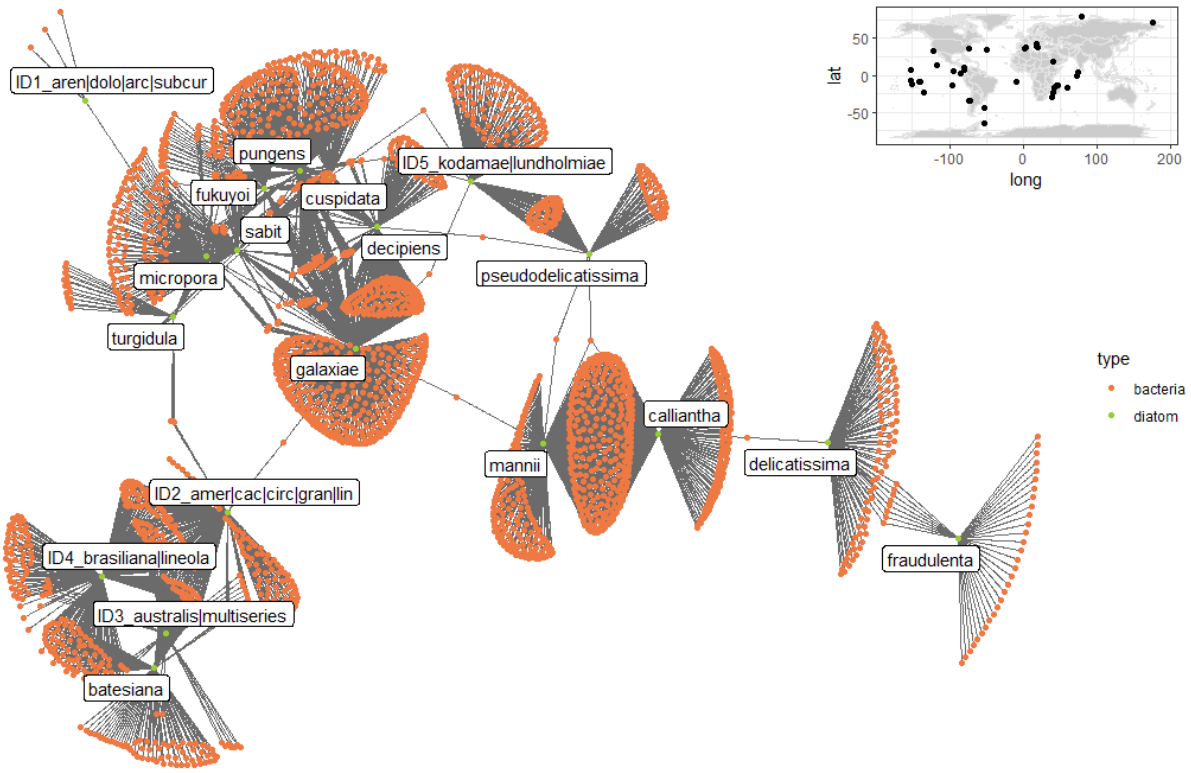

Figure 4.8. Correlation network visualizing significantly strong positive pairwise Spearman correlations between Pseudo-nitzschia species (green dots, and labelled) and particle attached prokaryotic OTUs (orange dots) across 40 sampling stations (top-right). A total of 19 Pseudo-nitzschia species ID and a 2355 prokaryotic OTUs were used to build the network. Both diatoms and bacteria belonged to the 20-180 µm size class.

The degree of each *Pseudo-nitzschia*, i.e., the number of edges of each node, has been calculated for both networks and is shown in Table 4.3.

Table 4.3. Degree per each Pseudo-nitzschia species ID as retrieved from both FL-based and PA-based networks.

Pseudo-nitzschia Species ID	FL	PA
batesiana	278	165
calliantha	/	396
cuspidata	91	48
decipiens	140	230
delicatissima	192	91

fraudulenta	47	38
fukuyoi	58	142
galaxiae	310	510
ID1_*aren*| *dolo*| *arc*| *subcur*	/	4
ID2_*amer*| *cac*| *circ*| *gran*| *lin*	300	246
ID3_*australis*| *multiseries*	41	89
ID4_*brasiliana*| *lineola*	637	290
ID5_*kodamae*| *lundholmiae*	360	159
mannii	167	438
micropora	53	91
pseudodelicatissima	304	84
pungens	175	298
sabit	245	143
turgidula	51	35

The degree of each *Pseudo-nitzschia* species for both networks has been plotted together with the mean relative abundance and the occupancy of each species ID (Fig. 4.9.). The mean relative abundance indicates the average value of the relative contribution of a species ID to the total abundance of all *Pseudo-nitzschia* species in each community, while the occupancy is defined as the absolute number of stations in which each species ID occurs. For both networks it seems that, while an expected link between mean relative abundance and occupancy existed, this does not explain the level of degree of each species ID.

The two networks are different in terms of absolute number of prokaryotic OTUs retrieved, with the PA-based network holding a higher number of bacteria. ID5 (*P. brasiliana*|*P. lineola*) and *P. calliantha* held the highest levels of degree for the FL- and the PA-based networks, respectively.

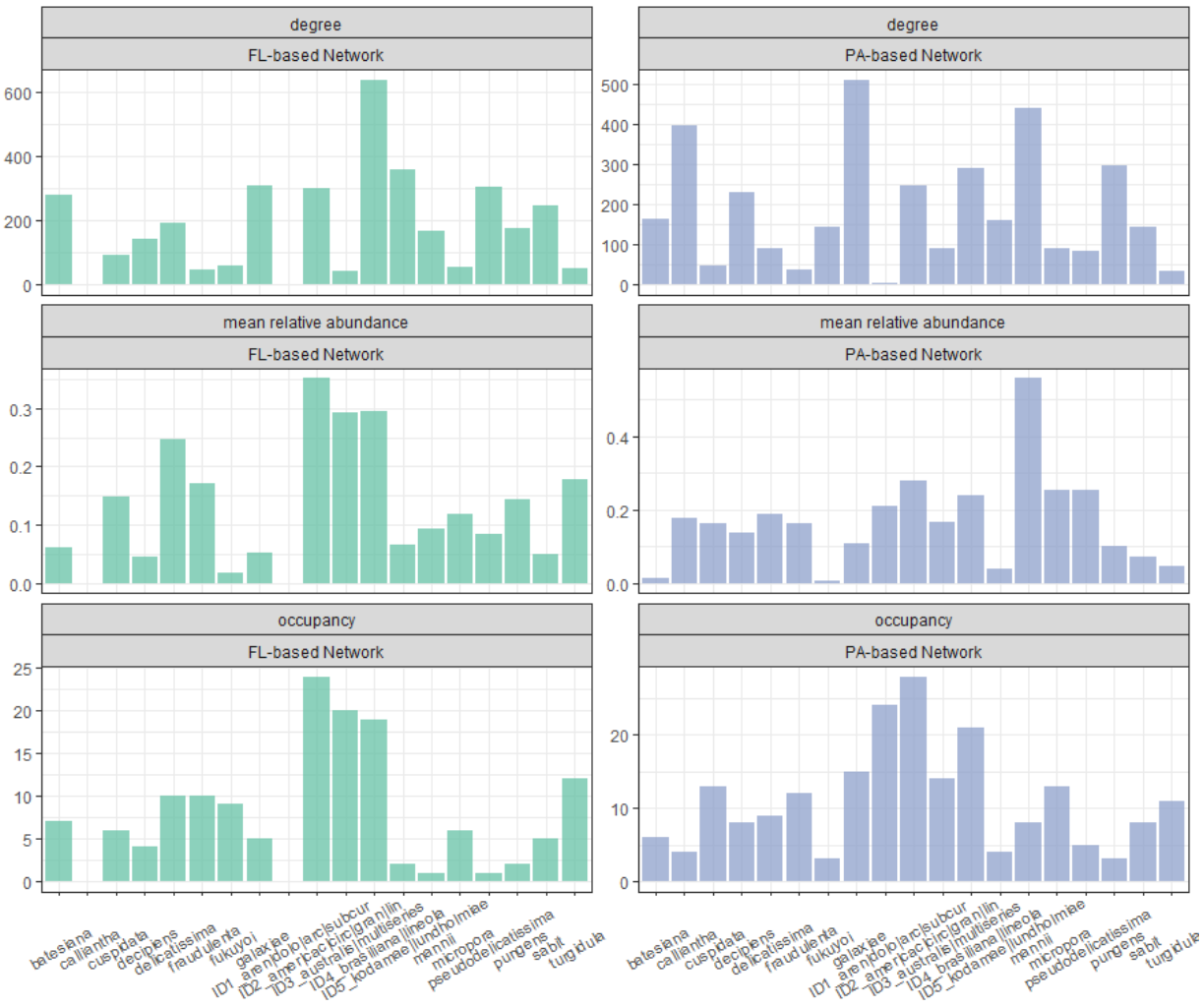

Figure 4.9. Degree, mean relative abundance and occupancy of each Pseudo-nitzschia species ID as retrieved from both FL-based (green) and PA-based (violet) networks.

In order measure the level of species-specificity of each bacterial community, I visualized the percentage of bacteria uniquely associated with a species, relatively to the total degree of that species. Results are shown in Fig. 4.10.

The PA-based network showed a high species-specificity for 5 species: *P. galaxiae*, *P. fraudulenta*, *P. cuspidata*, *P. delicatissima* and *P. micropora*, together with the group ID5 (*P. kodamae*|*lundholmiae*). The latter three showed high species-specificity of the bacteria associated also in the FL-based network, together with *P. turgidula*, *P. pseudodelicatissima*, *P. fukuyoi* and the species group ID3 (*P. australis*|*multiseries*). These results overall confirm what has been observed through experimental studies, i.e., that different bacterial communities inhabit the

phycosphere of different *Pseudo-nitzschia* species (*Guannel et al., 2011*; *Sison-Mangus et al., 2014*; *Lelong et al., 2014*).

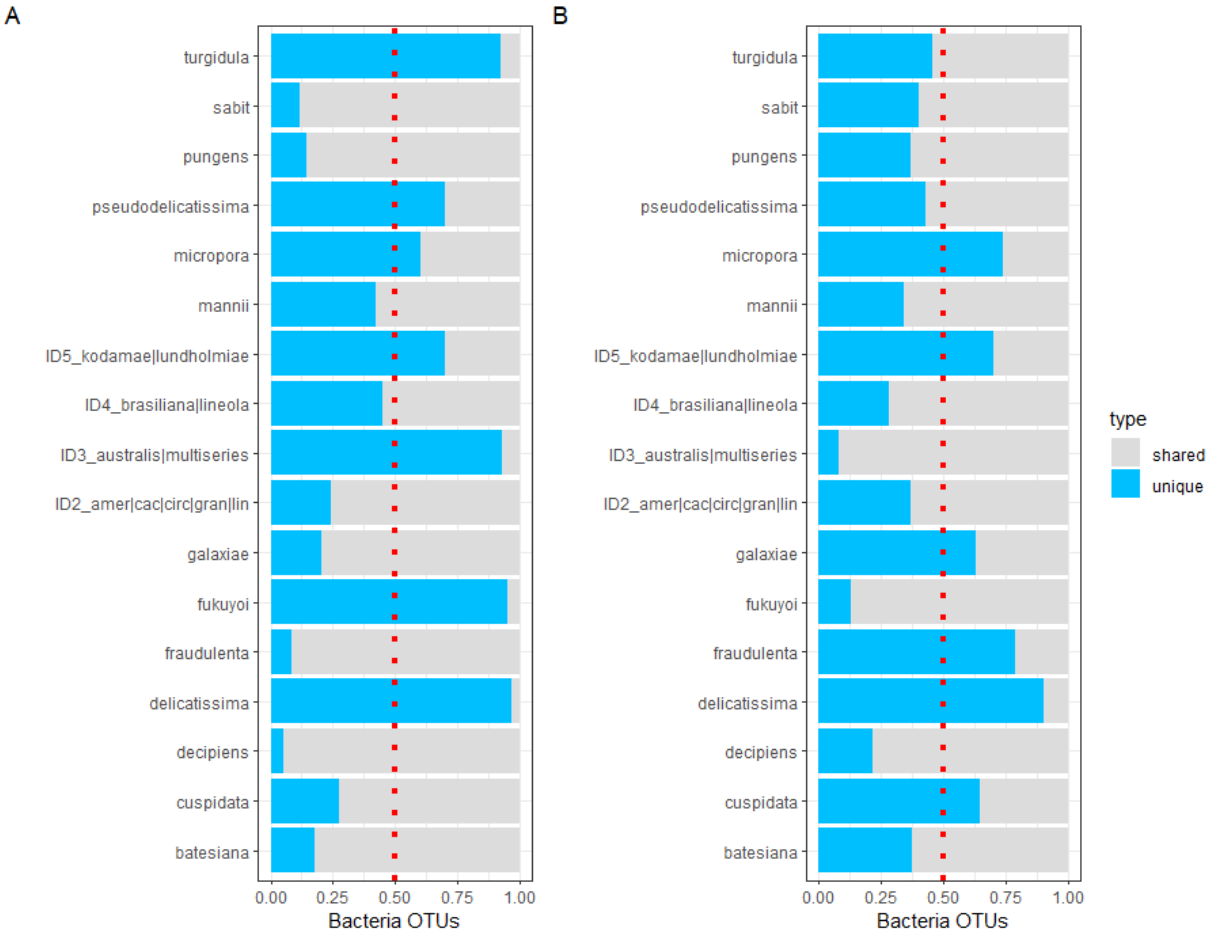

Figure 4.10. Bar plots showing the percentage of prokaryotic OTUs uniquely linked to each Pseudo-nitzschia species ID the A) FL-based and B) PA-based networks. Only Pseudo-nitzschia species present in both networks are shown.

Although a high level of species-specificity, bacteria that strongly co-vary with *Pseudo-nitzschia* species in both networks did not show the same feature when looking at their taxonomic annotation at the phylum level (Fig. 4.11); the unicity of the prokaryotic community of each diatom species seems thus to rely to distinct bacteria species or strains, a differentiation not detectable at higher taxonomical ranks. The results for lower ranks, e.g., family and genus (not shown), displayed

overall the same lack of species-specificity found through the analysis at the phylum level.

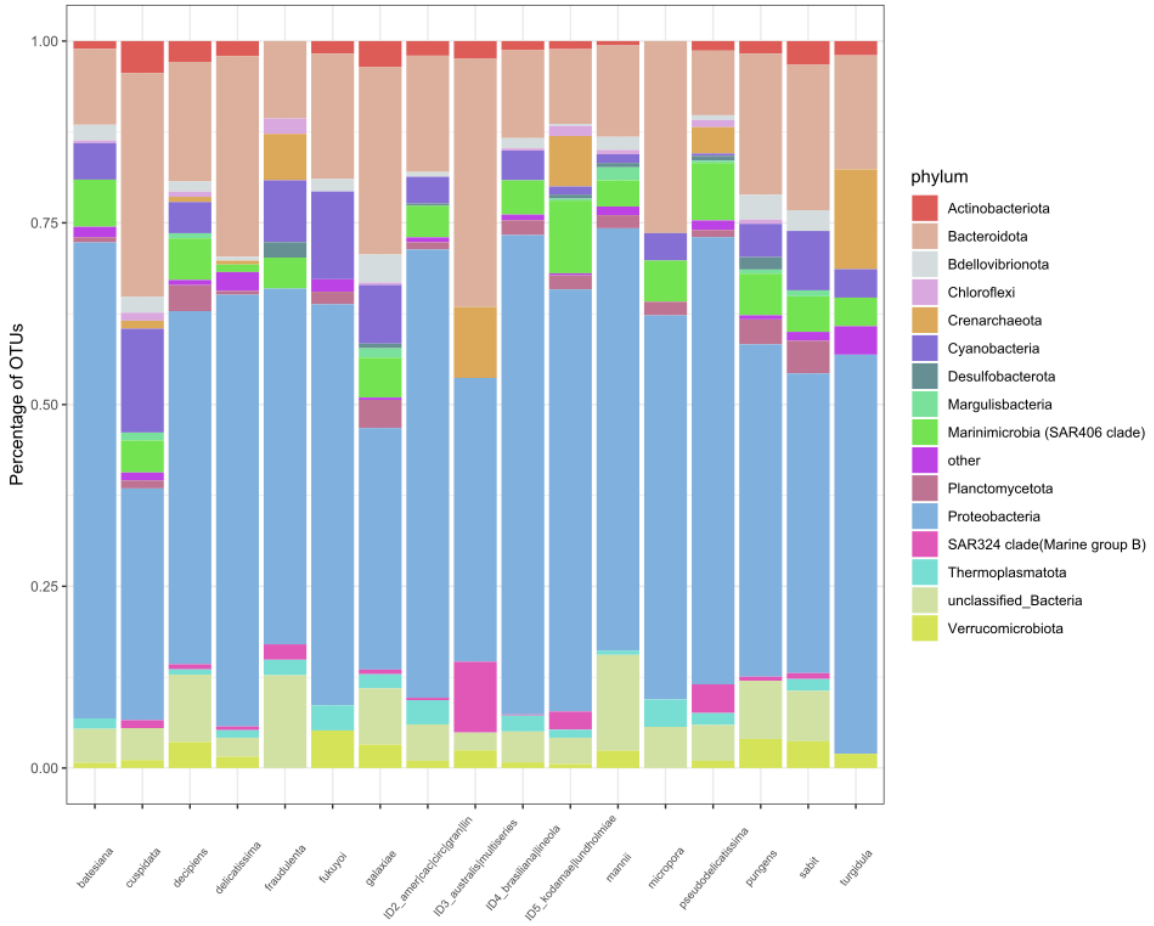

Figure 4.11. Bar plot indicating the taxonomical assignation at the phylum level for each Pseudo-nitzschia species for both FL- and PA-based networks. "Other" indicates phyla represented by less than 6 OTUs. Only Pseudo-nitzschia species present in both networks are shown.

Although the increasing understanding, corroborated by the present results, that different *Pseudo-nitzschia* species host distinct bacterial communities, whether there is a difference between the algae-attached and free-living bacteria remains uncertain, with studies indicating a lack of significant distinction between the two types of prokaryotic communities (*Guannel et al., 2011*), and others pointing to the existence of marker taxa as indicators of one or the other lifestyle (*Trano, 2021*). When looking at prokaryotic OTUs occurring in both FL and PA-based networks (Fig. 4.12), it emerged that the fraction of bacterial OTUs that occurred in both

networks was by far the smallest for each species, thus pointing to a difference between FL and PA two communities.

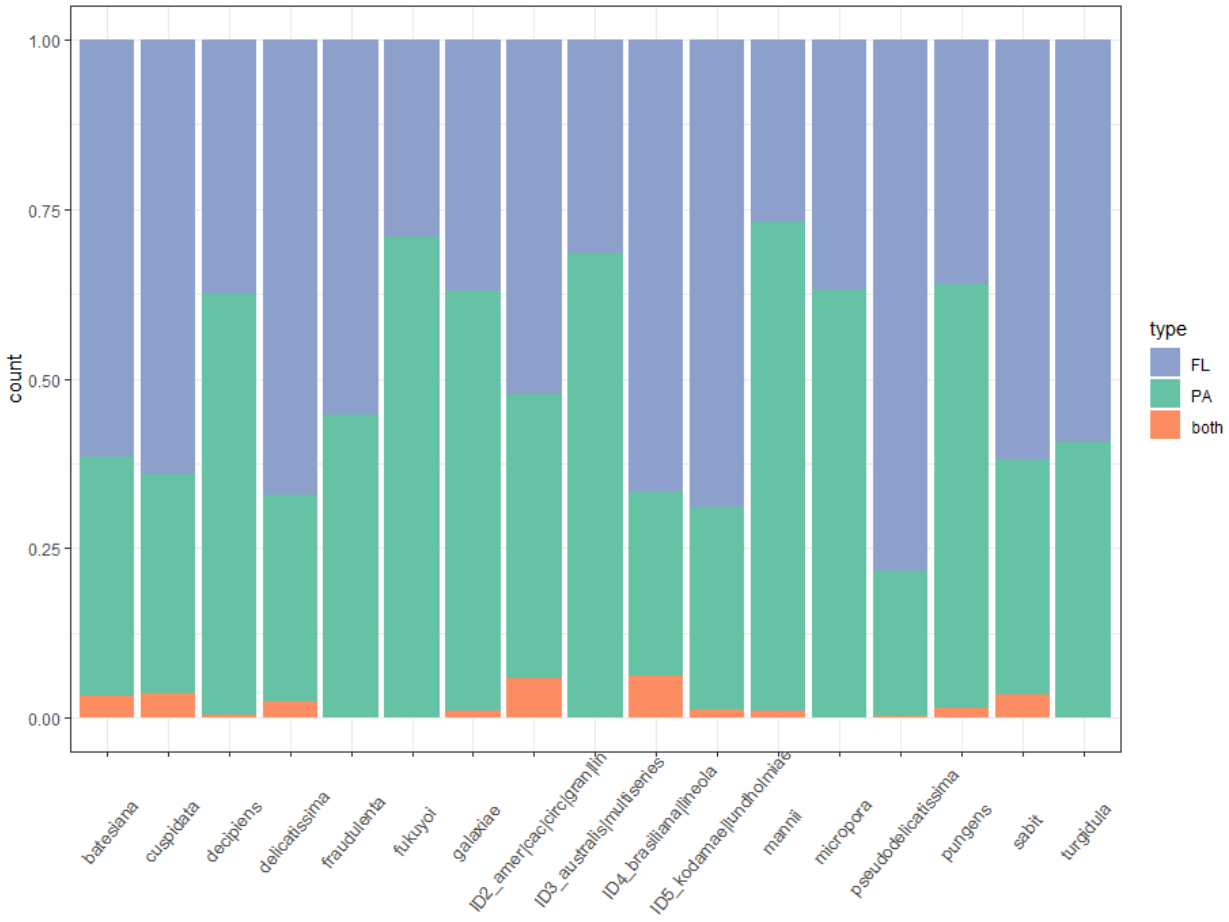

Figure 4.12. Bar plot indicating the relative percentage of bacteria co-variating with Pseudo-nitzschia species belonging exclusively to the free-living (FL, violet) or the particle-attached (PA, green) fraction, or occurring in both correlation networks (both, orange). Only Pseudo-nitzschia species present in both networks are shown.

The exploration of the taxonomic information of the three identified groups of bacteria, i.e., FL, PA, and the ones occurring in both networks (indicated as "both"), allowed a better characterization of prokaryotic communities that co-varied with *Pseudo-nitzschia* species. The Venn diagrams in Fig. 4.13 show that, while the bacteria occurring in both networks were mainly a subset of the other two groups, the differences between exclusively free-living and only-attached bacteria were reflected at higher taxonomic ranks, from the genus to the phylum level.

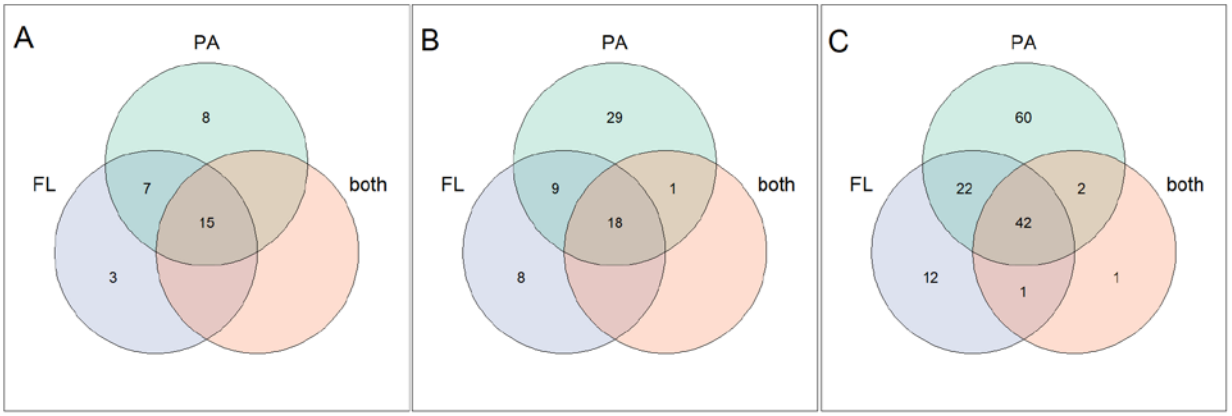

Figure 4.13. Venn diagrams showing the overlap between FL-only, PA-only and bacteria occurring in both networks across three taxonomical ranks: A) phylum, B) family, and C) genus.

In particular, 64 genera were shared by FL and PA bacteria, but the first group was also represented by 12 genera unique to this category, and the latter held 60 genera that were uniquely found in the large size fraction. Similarly, 29 families were solely found in PA bacteria and 8 were unique to FL bacteria, while 27 families were found for both groups. This difference also held at the highest taxonomic rank: 22 phyla were in common between PA and FL bacteria, 8 were specific to the PA and 3 to the FL category.

While all the prokaryotic OTUs were annotated at the phylum level, an increasing percentage of them were labelled as "unknown" in the dataset when moving towards lower taxonomic ranks, exceeding the 50% when looking at the genus level in the FL data set (Fig. 4.14 A). Free-living bacteria showed overall a higher amount of "unknown" compared to the particle-attached ones. Proteobacteria and Bacteroidota were the main represented phyla in both PA and FL microbial fractions, in agreement with what has been observed in marine systems (*Yung et al., 2016*) and for diatoms in particular (*Amin et al., 2012*; *Behringer et al., 2018*). The presence of these phyla was also a constant feature of bloom-associated dynamics, regardless the functional phytoplanktonic group involved in the bloom event (*Buchan et al., 2014*). Being specialized in degrading distinct types of organic matter produced by phytoplankton, the succession and the abundance oscillations of

these bacteria are driven and promoted by diatom blooms, with Proteobacteria, by transporting and using low molecular weight organic matter, acting in the first part of the algae proliferation, and Bacteroidota often contributing to degrade algae particles during the end of the bloom (*van Tol et al., 2017*).

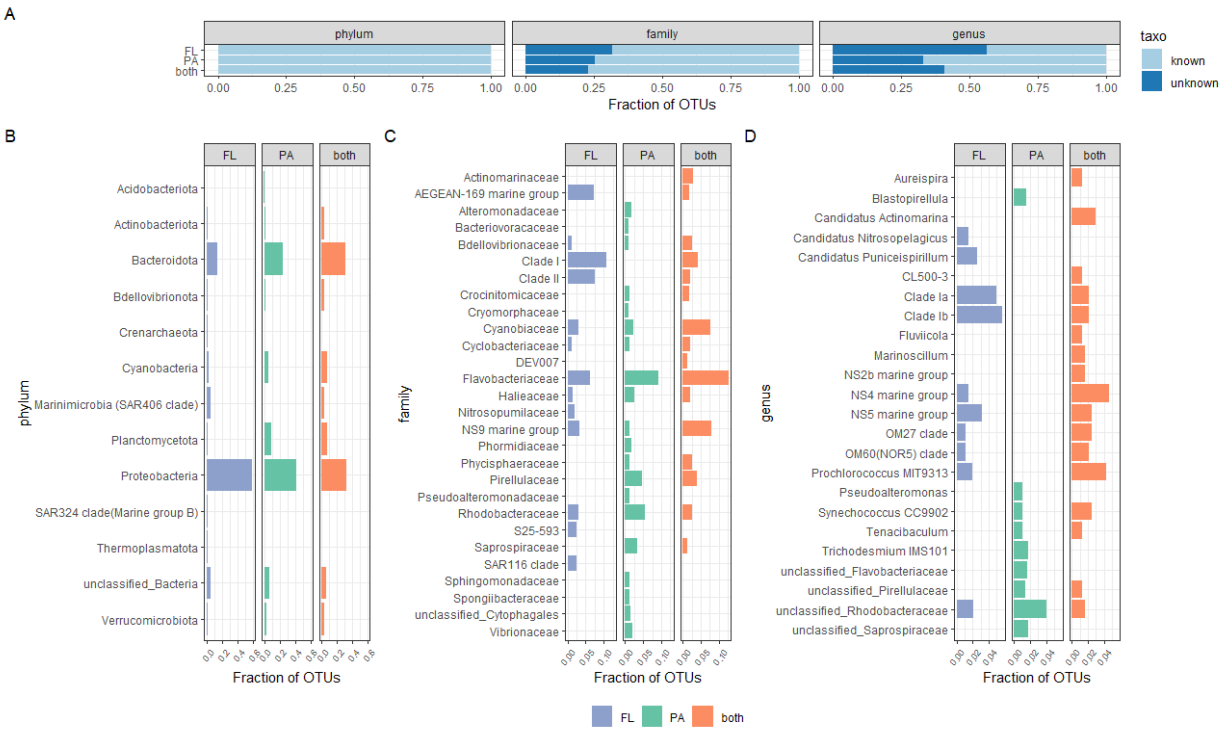

Figure 4.14. A) bar plot showing the relative proportion of unknown and annotated taxa at phylum, family and genus level. The relative abundance of OTUs belonging to the FL (violet) or PA (green) category as well as of bacteria occurring in both networks (orange) is showed by bar plots according to the taxonomical annotation of bacteria at B) phylum, C) family and D) genus level.

Flavobacteriaceae was the most abundant family for both the FL and PA fractions. These organisms, whose striking diversity is arranged in 100 different genera, are ubiquitous members of the global ocean, where they occupy different habitats and environments, being highly abundant in coastal waters and during phytoplankton blooms (*Alonso et al., 2007*; *Buchan et al., 2014*). Free-living or attached, Flavobacteriaceae are thought to play a key role in the degradation of complex polymeric organic matter (*Kirchman, 2002*; *Alonso-Saéz et al., 2007*). However, some members of this family have been shown to negatively impact diatoms by inhibiting their mitotic divisions (*Paul and Pohnert, 2011*; *van Tol et al., 2017*).

While a clear separation between taxa specialized to one of the other lifestyle (FL or PA) was impossible to detect when looking at bacteria from a high taxonomical rank, the investigation at family and genus levels allowed a better understanding of the differences between the prokaryotic OTUs retrieved from the two diatom-bacteria networks. The presence of significantly different FL and PA bacterial communities associated with diatoms has been recently assessed in a comprehensive study that used both morphological and molecular data from culture as well as *in situ* observations (*Trano, 2021*). The author provided a list of prokaryotic taxa that can be defined as indicators of one of the two categories of bacteria. Among them, the families of Pirellulaceae, Saprospiraceae and Sphingomonadaceae were defined as PA-specific, while SAR116 was labeled as a FL-specific taxon; my results confirmed what has been proposed, although the same concordance was not found for other taxa, including Rhodobacteraceae and Cyclobacteriaceae, proposed as PA-specific and found in both networks in my analysis. Moreover, Alteromonadaceae was more abundant in the PA fraction in the present results, although it has been proposed as FL-specific taxa (*Trano, 2021*).

At the genus level I found a clear separation between FL and PA. Genera found in FL-based network and never abundant in the PA network included members of *Candidatus Nitrosopelagicus* and *Candidatus Puniceispirillum*, together with Clade Ia and Clade Ib, NS4 and NS5 marine groups, and *Procholorococcus* MIT9313. PA-specific bacteria included *Blastopirellula*, *Pseudoalteromonas*, *Trichodesmium* IMS101, together with unclassified members of Flavobacteriaceae and Saprospiraceae. Bacteria that were found positively strongly correlated with diatoms in both networks displayed a wider range of genera, that included both the FL and the PA with a few taxa unique to this category, i.e., *Aureispira*, CL500-3, *Fluviicola*, *Marinoscillum* and NS2b marine group. This analysis suggests the presence of specialists and generalist bacteria. The first were preferentially associated with diatoms in one or the other size fraction, being thus adapted to the FL or PA lifestyle. Generalist bacteria are members of taxa that occurred in both

FL and PA fractions, but whose characteristics allow them to better exploit resources both when inhabiting the water that surrounds algae cells as well as when attached to diatoms' cellular surface.

4.4 Conclusion

Pseudo-nitzschia is a widespread genus, whose species represent an important component of diatom communities worldwide. The analysis presented in this chapter highlighted once more the ubiquitous nature of *Pseudo-nitzschia* species. However, the reference database used did not cover the whole diversity of the highly specious *Pseudo-nitzschia* genus, and the used metabarcoding marker, i.e., the V4 region of the 18S rDNA, lacked the necessary resolution to individually identify several species, that were indeed pulled together into species groups. Moreover, the biogeographic assessment here provided cannot be considered as complete, as we referred to samples often collected only once in a given location. Each point did not represent one single locality but it is the result of the combination of all environmental and biotic factors occurring at each station. It would thus be more correct to refer to different oceanographic conditions than regions or localities. Patterns of species occurrence assessed using several projects included in the comprehensive UniEuk data set showed how many species displayed a temperate-tropical distribution, while none was polar-endemic. Species occurring at the poles are thus also found elsewhere, and often represent cosmopolitan organisms tolerant of a wide range of temperature, salinity and nutrient concentrations. The integration of information on relative abundance across *Tara* Oceans and *Tara* Oceans Polar Circle sampling stations allowed a deeper understanding of the main biogeographical patterns: although detected in many regions across the global ocean, each species thrived and dominated *Pseudo-nitzschia* communities only in specific oceanographic conditions. Therefore, occurring everywhere does not mean to be abundant everywhere, suggesting that

Pseudo-nitzschia species can be locally rare taxa of a community capable to become dominant whether the environmental abiotic and biotic factors positively promote their growth. Biotic factors involved in *Pseudo-nitzschia* ecology are represented by the interactions with other organisms. I here studied correlation networks representing the positive co-variation of both free living (FL) and putatively particle attached (PA) bacteria with *Pseudo-nitzschia* species, that showed overall a high level of species-specific associations, in agreement with what was already experimentally demonstrated (*Guannel et al., 2011*; *Sison-Mangus et al., 2014*; *Lelong et al., 2014*). This species-specificity is mainly found at the OTU level; when looking at high rank taxonomic assignation, there is no evidence of a clear set of prokaryotic taxa that preferentially co-vary with one or another *Pseudo-nitzschia* species. Differences are instead found when comparing free-living and particle-attached bacterial communities, that overall show a very limited overlap at the OTU level. Both the FL and PA categories are dominated by phyla known to be the most abundant in marine systems (Proteobacteria, Bacteroidota), and the same is true for the family Flavobacteriaceae. While moving from a high (phylum) to a low (genus) taxonomic level, a clear separation between FL and PA communities emerges: some families seem to be specific to one or the other category, often matching taxa recently proposed as indicators of one or the other life-style (*Trano, 2021*). At the genus level the separation is obvious: a set of genera is clearly FL-specific, while others are unique to PA bacteria. Prokaryotic OTUs that occur in both FL- and PA-based networks have members occurring in both the FL- or PA-specific genera. Moreover, some genera are also unique to these generalist bacteria. Although each relationship detected between *Pseudo-nitzschia* and bacteria is to be interpreted only as putative association, these results are promising, and encourage the validation through experimental studies, also necessary to understand the nature of the interactions. For example, particle-attached bacteria are thought to often live in commensalism or mutualism with microalgae, and the nature of this interaction is strictly linked to the role of exchanged molecules (*Amin et al., 2015*; *Seymour et al., 2017*). Phytoplankton-bacteria

interactions are also strongly shaped by seasonality, whose role in determining both *Pseudo-nitzschia* biogeography and correlation networks has been overlooked in the present chapter. In this context, the integration of time-series data could provide additional information to understand the dynamics that rule the complex interaction of diatoms and bacteria.

Chapter V

Functional exploration of *Pseudo-nitzschia*: The case of sexual reproduction genes

5.1 Introduction

Diatoms have a diplontic life cycle where the diploid cells divide mitotically across long vegetative periods interspersed with short phases of sexual reproduction (*Chepurnov et al., 2005*). Sexual reproduction occurs with different strategies between centric and pennate diatoms, the first showing homothallic reproduction, with eggs and sperm produced from a clonal strain, the second, including *Pseudo-nitzschia* species, characterized by heterothallism, with gametes of strains of opposite mating types (*Montresor et al., 2016*).

As discussed in paragraphs 1.1 and 1.3.2, sexual reproduction is a way to restore the average size of diatom cells, whose rigid silica walls impose cell reduction during mitosis, and it occurs only when individuals reach a species-specific size threshold. Moreover, as observed for *Pseudo-nitzschia multistriata* (*Scalco et al., 2014*), a threshold on cell concentration must also be reached to allow sexual reproduction to start. A high cell density results in the spatial proximity of cells, indispensable for a successful encounter and chemical communication between the two opposite mating types; the role of chemical cues in this context has indeed been shown for benthic (*Sato et al., 2011*; *Gillard et al., 2013; Bilcke et al., 2021*) as well as for pelagic diatoms (*Basu et al., 2017*). In addition to length and/or size concentration, heterothallic pennate diatoms undergo sexual reproduction only under stable and non-stressful condition, being an exception among protists, where sex is often promoted by unfavourable conditions such as starvation,

oxidative stress or sharp fluctuations of environmental factors (*Nedelcu et al., 2004*; *Goodenough et al., 2007*; *Montresor et al., 2016*). Once sexual reproduction has been induced, meiotic divisions lead to the formation of gametes; haploid gametes of opposite mating types then conjugate, producing two expandable diploid zygotes that ultimately develop into auxospores (*Davidovich and Bates, 1998*). Within each auxospore, an initial cell of maximum size is synthesized, thanks to the absence of the rigid siliceous cell wall, restoring the vegetative phase of the cycle. Morphologically indistinguishable, the two mating types are conventionally referred to as mating type plus (MT+) and minus (MT-): the first indicates the strain able to produce two gametes that move towards the sessile gametes held in the MT- gametangium, i.e., the one that carries the auxospore (*Montresor et al., 2016*).

But what happens at the molecular level during sexual reproduction events?

Experimental studies using genomic and transcriptomic approaches demonstrated that the cellular release of chemical clues during *P. multistriata* sexual reproduction occurs in its early phases and induces changes in gene expression in both MT+ and MT- (*Basu et al., 2017*). For instance, regulatory proteins of cell cycles (i.e. cyclins) as well as genes involved in nutrients uptake and transport, are down-regulated in both MT+ and MT- once *P. multistriata* cells undergo sexual reproduction. Moreover, observations and flow cytometry analyses demonstrated that cell growth stopped during sexual reproduction (*Scalco et al., 2014*; *Basu et al., 2017*). These results were recently confirmed by a study (*Annunziata et al., 2022*) that reported the down-regulation of genes related to major metabolic functions during *P. multistriata* sexual phase, also adding photosynthesis-related genes to the list and finding changes in the transcription levels of genes involved in lipid metabolism. Furthermore, through analyses built on numerical models, the authors proposed that the confirmed growth arrest in parental cells involved in sexual reproduction causes an alteration in the parent-sibling equilibrium in favour of the new generations.

The above-mentioned results, together with the most recent insight, not only point to the fundamental role of sexual reproduction in the life cycle of diatoms, whose occurrence profoundly affects the transcriptomic profile of cells, but also highlight the importance of life cycle traits in shaping population dynamics that ultimately determine the ecological structure of communities. However, the study of sexual reproduction in field populations remains challenging, especially because of the scarce and unpredictable nature of sexual reproduction events, together with the difficulty in identifying sexual stages through morphological approaches (*Mann, 1988*). The improvement and advancements in genomic and transcriptomic approaches, together with the availability of meta-omic data, and especially metatranscriptomics, offers a unique opportunity to investigate processes related to organismal physiology, metabolism, as well as life cycle traits. First, a comparative genomic study coupled with RNA-sequencing approaches identified a conserved set of 42 genes that played a role in diatom meiosis, out of which five were exclusively involved in the meiotic phase in *Seminavis robusta* and *P. multistriata* (*Patil et al., 2015*). This list was successively refined and integrated with the addition of a new set of sex-induced genes (SIGs) identified through RNA-sequencing approaches using the centric diatom *Skeletonema marinoi* (*Ferrante et al., 2019*). More recently, a gene family-based comparative analysis allowed the identification of a set of 10 diatom gene families with a significantly high sex-specificity (*Bilcke, 2021*); a further refinement of this list led to the final selection of a panel of 5 genes that can be used as molecular markers to detect sexual reproduction events in natural populations through the investigation of metagenomic and metatranscriptomic data (*Borgonuovo et al., in prep.*).

In this context, in the present chapter I will identify and characterize the gene expression of the recently proposed molecular markers for sexual reproduction in diatoms using the *Tara* Oceans and *Tara* Oceans Polar Circle metatranscriptomic data set, with the integration of metabarcoding and environmental data and with a special focus on *Pseudo-nitzschia* species. In order to track the occurrence of

sexual reproduction in diatoms at sea, I will first exploit *Tara* meta-omic data set selecting only genes phylogenetically close to the ones whose up-regulation during sexual reproduction was experimentally proved. I will then localize the presence of each proposed marker at global scale and then use patterns of gene co-expression as an indication of the occurrence of sexual events at sea. Metabarcoding information and environmental data will also be integrated to support the evidences provided by the functional exploration.

5.2 Materials and Methods

5.2.1 Selection of sexual reproduction markers

The selection of sexual reproduction markers has been performed by Dr. Rossella Annunziata (Stazione Zoologica Anton Dohrn) and Dr. Gust Bilcke (Gent University). A summary of the procedure that led to the final selection is described below. More information can be found in *Bilcke (2021)* and will be published in *Borgonuovo et al. (in prep.)*. Data from the species *Cylindrotheca closterium* (unpublished) were generated by Dr Sien Audoor (Gent University).

Transcriptomic data from different sexual reproduction experiments of four species were compared. In particular, the four species consisted of one centric diatom, *Skeletonema marinoi*, and three pennate diatoms, i.e., the benthic *Seminavis robusta*, the meroplanktonic *Cylindrotheca closterium* and the pelagic *Pseudo-nitzschia multistriata*. For each species, transcriptomic data were compared at four different mating stages, namely sex pheromone signalling (S), gametangia (P, pre-gametic), gametes/zygotes (GZ) and auxospore (A). The quality-filtered RNA-sequencing reads were mapped to their respective transcriptomes and, following normalization procedures on all libraries, differential expression (DE) tests were performed, contrasting the response in sexual and control samples for each

species. This analysis, followed by statistical validations, allowed to define genes that were up- or down-regulated at the different time points. Once the DE results for the four species were obtained, homology between species was assessed depending on shared homologous gene families; in particular, the TribeMCL approach (*Enright et al., 2002*) was used for *S. robusta* and *P. multistriata*, with functional information retrieved from PLAZA Diatoms 1.0 (https://bioinformatics.psb.ugent.be/plaza/versions/plaza_diatoms_01/; *Osuna-Cruz et al., 2020*). The same database was also used as the reference source to allow the functional annotation of *S. marinoi* and *C. closterium* genes using the tool TRAPID 2.0 (*Bucchini et al., 2021*), as described in *Bilcke (2021)*. The obtained gene families were then analysed with the purpose of investigating whether they contained genes both sufficiently conserved and sex-specific to be labelled as diatom sexual markers. In particular, three conditions were set as necessary in order to select the final set of gene families:

1. The gene family had to be diatom specific;
2. The gene family had to be present in at least three out of the four diatom species considered;
3. All the expressed genes of a gene family had to be significantly up-regulated in at least one mating stage in all the four species.

Finally, the top 5 gene families that met the above-mentioned requirements and showed the highest average cross-species fold changes were considered, leading to a final selection of five marker gene sets, named M1 to M5.

5.2.2 Notes on markers

Genes of the M1 family are characterized by a carboxyterminal Tubby-like domain, a domain shown to be involved in the coordination of multiple signalling pathways in studies on model organisms belonging to both plant and animal kingdoms (*Mukhopadhyay and Jackson, 2011*). The M2 family contains a concanavalin A-like lectin/glucanase protein domain. Lectins and glucanases are highly conserved

binding proteins found in prokaryotes as well as humans as mediators of cellular recognition (*Lis and Sharon, 1998*). They are able to undergo reversible bindings to complex carbohydrates and all the family members contain a concavalin A-like domain, involved in sexual reproduction with a role in gamete-gamete recognition in red (*Shim et al., 2012*) and brown (*Bolwell et al., 1979*) algae, as well as in mice (*Clark, 2011*), and found in hormone-binding globulins involved in the transport of human sex steroids in blood (*Grishkovskaya et al., 2000*). M3 corresponds to a gene family whose members have already been proposed as molecular markers for sexual reproduction in diatoms, belonging to the identified panel of sex induced genes (SIGs; *Ferrante et al., 2019*), and in particular corresponding to SIG7. These genes contain a homologous-pairing protein 2 (hop2) domain, involved in chromosome pairing and meiotic recombination. However, hop2 is presumed to be absent in diatoms (*Patil et al., 2015*; *Ferrante et al., 2019*), and whether SIG7 genes can be considered as homologs to hop2 proteins is still to be clarified (*Bilcke, 2021*). It is noteworthy that, besides the case of SIG7, the other proposed SIGs, although highly expressed during sexual reproduction of the considered species, are not exclusively upregulated in this phase and thus do not fulfil the requirements to be sex-specific marker genes (*Bilcke, 2021*). The M4 family does not contain any recognized protein domain, while M5 holds a tetratricopeptide domain, known to be involved in protein-protein interaction and molecular recognition (*Blatch, and Lässle, 1999*; *Perez-Riba and Itzhaki, 2019*). *P. multistriata* homologs occurring in each of the above-mentioned gene families were significantly up-regulated during the gamete/zygote mating stage. In addition to the newly proposed markers, the meiosis-specific SPO11-2 gene was included in the analysis. In particular, since SPO11-2 up-regulation during sexual reproduction of *P. multistriata* and *S. robusta* was recently experimentally demonstrated (*Patil et al., 2015*), the information on presence and abundance of this gene in metagenomic and metatranscriptomic data across *Tara* Oceans and *Tara* Oceans Polar Circle sampling stations was used to filter, confirm and corroborate the results obtained with the M1-M5 gene

families. Therefore, the final panel of considered markers for the analysis presented in this chapter contained 6 gene families: M1-M5 and SPO11-2.

5.2.3 Meta-omic exploration of diatom sexual reproduction markers

The pipeline presented here, included in the following sections (5.2.3.1. to 5.2.3.3.), has been developed and executed in collaboration with Dr Camilla Borgonuovo (Stazione Zoologica Anton Dohrn). The pipeline is similar to the one already described in Chapter II (sections 3.2.4.1. – 3.2.4.2.).

5.2.3.1 Sequences search

Multiple protein sequence alignments for each gene family were carried out using MUSCLE algorithm (*Edgar, 2004*) implemented in the software MEGAX, version 11.0.8 (*Kumar et al., 2018*), and then transformed into Profiles Hidden Markov Models (HMM) using the command hmmbuild as part of the HMMer package, version 3.1b2 (*Finn et al., 2011*). The HMM profiles were used as queries for sequence similarity searches using the HMMER algorithm against the MATOU-v2 dataset (Marine Atlas of *Tara* Oceans unigenes; *Carradec et al., 2018*; see paragraph 1.4.1.3). This catalogue was consulted through the Ocean Gene Atlas website (OGA; http://*Tara*-oceans.mio.osupytheas.fr/ocean-gene-atlas/; *Villar et al., 2018*) by leaving the default threshold on the e-value, i.e., the number of expected hits of similar quality that could be found just by chance; in particular, only sequences with an e-value lower than $1E^{-10}$ were selected. The exact number of sequences used to build HMM profiles for each gene and their taxonomic affiliation is provided in Table 5.1.

Table 5.1. Number of sequences used to build HMM profiles for each marker. Pmse: Pseudo-nitzschia multiseries; Pmu: Pseudo-nitzschia multistriata; Fcy: Fragilariopsis cylindrus; Sro: Seminavis robusta; Fso: Fistulifera solaris; Smo: Skeletonema marinoi; Ptri: Phaeodactylum tricornutum; Sya: Synedra acus; Ccl: Cylindrotheca closterium; Tho: Thalassiosira oceanica; Tpse: Thalassiosira pseudonana; Ccr: Cyclotella cryptica.

Marker	Pmse	Pmu	Fcy	Sro	Fso	Smo	Ptri	Sya	Ccl	Tho	Tpse	Ccr
M1	1	1	1	1	0	0	0	0	1	0	0	0
M2	1	1	1	1	3	0	1	0	1	0	0	0
M3	1	1	1	1	1	3	1	1	1	2	1	1
M4	1	1	1	2	0	0	0	0	1	0	0	0
M5	2	1	2	1	2	1	1	1	1	1	1	1
SPO11-2	1	1	1	1	2	1	1	2	1	1	1	1

Notwithstanding the use of annotated sequences to build the HMM profiles and the imposed threshold on the e-value, a considerable number of false positives is expected to be obtained by searching the gene catalogue. Therefore, the output hits were furtherly filtered according to their taxonomic assignation and bit score values. In particular, retrieved unigenes were grouped according to their phylum annotation into four categories: Bacillariophyta, Other Stramenopiles, Other Taxa, Unknown. Genes belonging to Other Taxa, retrieved only for M1 and SPO11-2, were excluded from the analysis of M1, and retained for SPO11-2 only if showing a bit score ≥ 100. Finally, all the retrieved hits for family markers 2, 3, 4 and 5 were retained for the subsequent analysis.

5.2.3.2 Phylogenetic analysis

The filtering procedure based on taxonomy annotation and bit score values allowed a first screening of the results. In order to select MATOU-v2 hits that more robustly represented homologs to the query sequences used to build the HMM profiles, we applied an approach that provided insight into the phylogenetic relationships between sequences.

First, for each gene family, the selected MATOU-v2 sequences were aligned with the diatom queries used to build the original HMM profiles. Multiple sequence alignments were then performed using the MUSCLE algorithm (*Edgar,* 2004) implemented in the software MEGAX, version 11.0.8 (*Kumar et al., 2018*) followed by automatic removal of poorly aligned positions with trimAl v1.4. (*Capella-Gutierrez et al., 2009*) (gap treshold -gt 0.1). Maximal likelihood phylogenetic trees with 1000 ultrafast bootstrap replicates were constructed using IQ-TREE v2.1.3 (*Nguyen et al., 2015*), with automatic ModelFinder substitution model selection. Trees were visualized and edited with FigTree version 1.4.3 (*Rambaut and Drummond, 2012*); the editing was relative to the layout of trees and gene IDs of sequences that were manually renamed with their taxonomic assignation. According to tree topology and bootstrap support values, a visual and manual selection of unigenes was independently performed for each phylogenetic tree. Finally, a filtering step based on consistency among trees was applied, with the rationale that to assess the occurrence of diatom sexual reproduction at a sampling site, the simultaneous expression of markers from ideally the same species must occur. Therefore, only MATOU-v2 unigenes that clustered in strongly supported clades with the same diatom species across the phylogenetic trees of all markers were retained for the subsequent analysis.

5.2.3.3 Final set of genes

Upon having obtained the catalogue of genes that had undergone bit score, taxonomy and phylogenetic-driven selection, we retrieved the information on the abundance of each MATOU-v2 gene in each sample of interest for both metagenomic and metatranscriptomic data. Unigene abundance is expressed as "percent of total genes per sample", that is gene read coverage in RPKM (Reads Per Kilobase covered per Million of mapped reads) divided by the sum of the total gene coverage for the sample, so that the abundance results are the fractions of homologs of all genes in the whole sample. Our assumption is that the selected

markers are all together exclusively involved in diatom reproduction; therefore, their simultaneous expression in a *Tara* station would constitute a strong indication that diatom sexual reproduction event is occurring in that sampled location. Therefore, after illustrating the geographical distribution of each marker separately, co-expression patterns were investigated using SPO11-2 and at least one other marker. Since we could not verify if SPO11-2 and the other marker genes belonged to the same species, this analysis has to be considered as qualitative.

5.2.3.4 Co-expression of markers

Sampling stations in which there was the simultaneous expression of SPO11-2 and at least one of the proposed sexual markers were identified and plotted in R using the packages *ggplot2* (Wickham, 2016) and *Scatterpie* (Guangchuang, 2020). Bar plots showing the number of markers in samples where SPO11-2 is co-expressed with at least another marker were constructed.

5.2.3.5 Metatranscriptomic and metabarcoding richness

First, patterns of richness were calculated for samples where SPO11-2 was co-expressed with at least another marker. The functional richness was indeed calculated as the total number of transcribed unigenes detected in a sample. As a consequence of the phylogenetics-driven gene selection (see section 5.2.3.2.), we considered these transcripts as belonging to *Pseudo-nitzschia* or *Fragilariopsis* species, as they were the only two taxa that clustered together with MATOU-v2 homologs across all markers. Therefore, metatranscriptomic richness has been compared with the number of distinct *Pseudo-nitzschia* and *Fragilariopsis* OTUs in a sample. This number, that reflects metabarcoding richness, was independently calculated using information from both the V4 and V9 regions of the 18S rDNA metabarcoding marker. In order to understand whether the emerging coherence between metatranscriptomic and metabarcoding richness was a specific signal from the two selected genera or the result of an overall greater richness of the

whole diatom community in the selected samples, the same metatranscriptomic richness was also compared with metabarcoding richness calculated over the amount of all diatom OTUs, thus including also genera other than *Pseudo-nitzschia* and *Fragilariopsis*.

5.2.3.6 Integrating relative abundance information from metabarcoding data

Relative abundance of *Pseudo-nitzschia* and *Fragilariopsis* OTUs was calculated over the total abundance of diatom OTUs in samples where SPO11-2 showed co-expression with at least another marker. A Venn diagram and an Upset plot were created to show the number of shared unigenes among stations and samples where the co-expression of SPO11-2 with at least another marker was matched by a dominance of *Pseudo-nitzschia* or *Fragilariopsis* OTUs (relative abundance > 50%). The used R packages were *venn* (*Dusa, 2020*) and *UpSetR* (*Gehlenborg, 2019*).

5.2.3.7 Chlorophyll a

With the aim of detecting putative bloom events in target samples, Chlorophyll *a* concentration was used as an indicator of phytoplankton biomass and plotted across all *Tara* Oceans and *Tara* Oceans Polar Circle surface stations.

5.3 Results and Discussion

5.3.1 Sequences search and raw hits selection

As shown in Table 5.1, a different number of sequences, ranging from a minimum of 5 to a maximum of 15, were used for each marker gene family to build the HMM profiles that were used to query the MATOU-v2 dataset (Marine Atlas of *Tara* Oceans Unigenes) through the Ocean Gene Atlas (OGA) website. The number of sequences retrieved for each marker from the MATOU-v2 database is shown in

Table 5.2, while their taxonomic assignation and the associated bit score values are shown in Fig. 5.1.

Table 5.2. Number of hits retrieved from MATOU-v2 through HMM search for each marker.

Marker	Number of retrieved hits	Number of selected hits
M1	308	214
M2	51	51
M3	170	170
M4	41	41
M5	312	312
SPO11-2	983	635

Although the selected gene families were diatom-specific (see section 5.2.1), a variable amount of non-diatom sequences was retrieved through the HMM profile search. The only marker that uniquely displayed diatom sequences (i.e., annotated as Bacillariophyta) was M4; a small number of non-diatom sequences was obtained for M2, M3 and M5, while this number increased for M1 and SPO11-2, where diatoms were the lowest (M1) or the second lowest (SPO11-2) represented group. Bit score values were always ≥ 50, a value usually considered as statistically significant for inferring homology in protein alignments (*Pearson, 2013*). All the retrieved hits of M2, M3 and M5 were kept for the subsequent analysis; genes not belonging to diatoms ("Other Taxa" in Fig. 5.1.) were excluded from the analysis of M1, and retained for SPO11-2 only if showing a bit score ≥ 100. The number of selected hits for each marker is reported in Table 5.2.

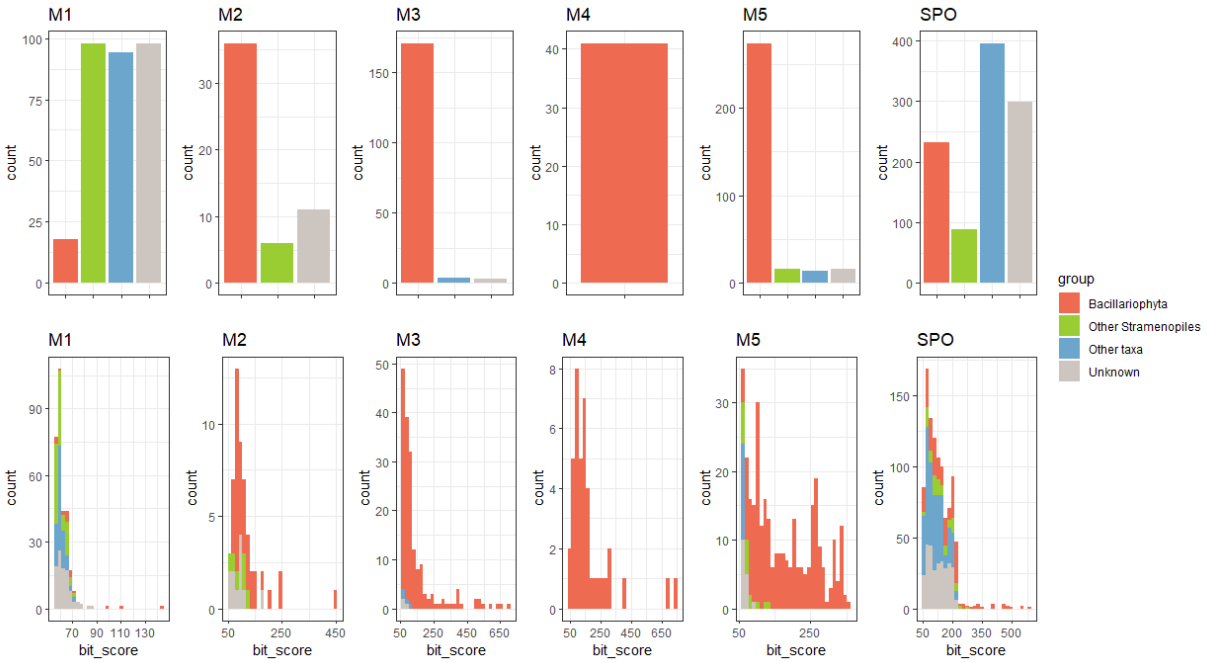

Figure 5.1. Numbers (top) and bit score values (bottom) of the hits retrieved for each marker through HMM search in MATOU-v2 database (Marine Atlas of Tara Oceans Unigenes). Colours of bars indicate the taxonomic annotation of sequences.

5.3.2 Phylogenetic analysis

Following raw hits selection, a phylogenetic tree containing the selected MATOU-v2 unigenes and the sequences used to construct HMM profiles (query sequences) was built for each marker gene family. The visual inspection of phylogenetic trees allowed to further select MATOU-v2 hits depending on their proximity to known diatom sequences and on bootstrap values. A brief description of each phylogenetic tree is found below.

M1

The HMM search of M1 genes resulted in 308 MATOU-v2 hits, out of which 214 were used for phylogenetic analysis (Table 5.2). The tree in Fig. 5.2. shows one strongly supported clade (bootstrap value: 100) where both *Pseudo-nitzschia* query sequences clustered together with three MATOU-v2 sequences out of which two were annotated at the family level (Bacillariaceae) and one was of unknown taxonomy. A second clade contained the *S. robusta* query and one unknown

Bacillariophyta sequence. The remaining 210 sequences clustered together in a large and strongly supported clade that did not hold any of the query sequences.

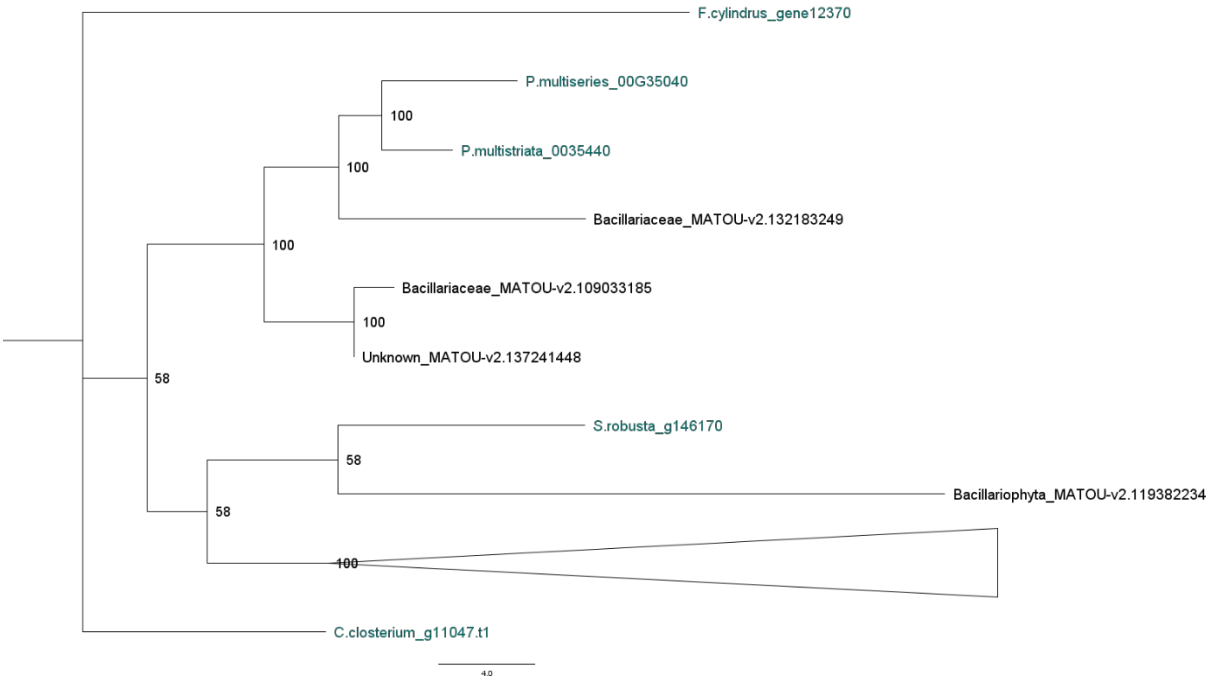

Figure 5.2. Maximum Likelihood (ML) tree of M1 genes. Numbers at the base of each node refer to bootstrap support after 1000 replicates. Colours refer to the origin of sequences: in black sequences retrieved from the MATOU-v2 database, in pine green diatom sequences. The triangle indicates collapsed branches.

M2

The HMM search of M2 genes resulted in 51 MATOU-v2 hits, all used for the phylogenetic analysis (Table 5.2).

The tree in Fig. 5.3. shows one clade composed by the query sequences of *F. solaris* and *P. tricornutum*, together with one MATOU-v2 unigene annotated as Bacillariophyceae. A second bigger clade contained the majority of the known query sequences. *F. cylindrus* query robustly clustered with two MATOU-v2 sequences annotated at the order level as Bacillariales. A poorly supported clade

included two MATOU-v2 unigenes. Another clade, whose internal nodes showed overall low bootstrap values, contained both the *Pseudo-nitzschia* queries used and 13 MATOU-v2 genes, out of which the majority was annotated at the order (Bacillariales) or family (Bacillariacae) level, with the exception of one hit ascribed to the genus *Pseudo-nitzschia*. Another *Pseudo-nitzschia* MATOU-v2 sequence was found in a separate and poorly supported clade. Queries of *S. robusta* and *C. closterium* clustered together in a strongly supported clade (bootstrap value: 72), while the majority of MATOU-v2 hits were found in clades (collapsed in Fig. 5.3.) where no known sequences occurred.

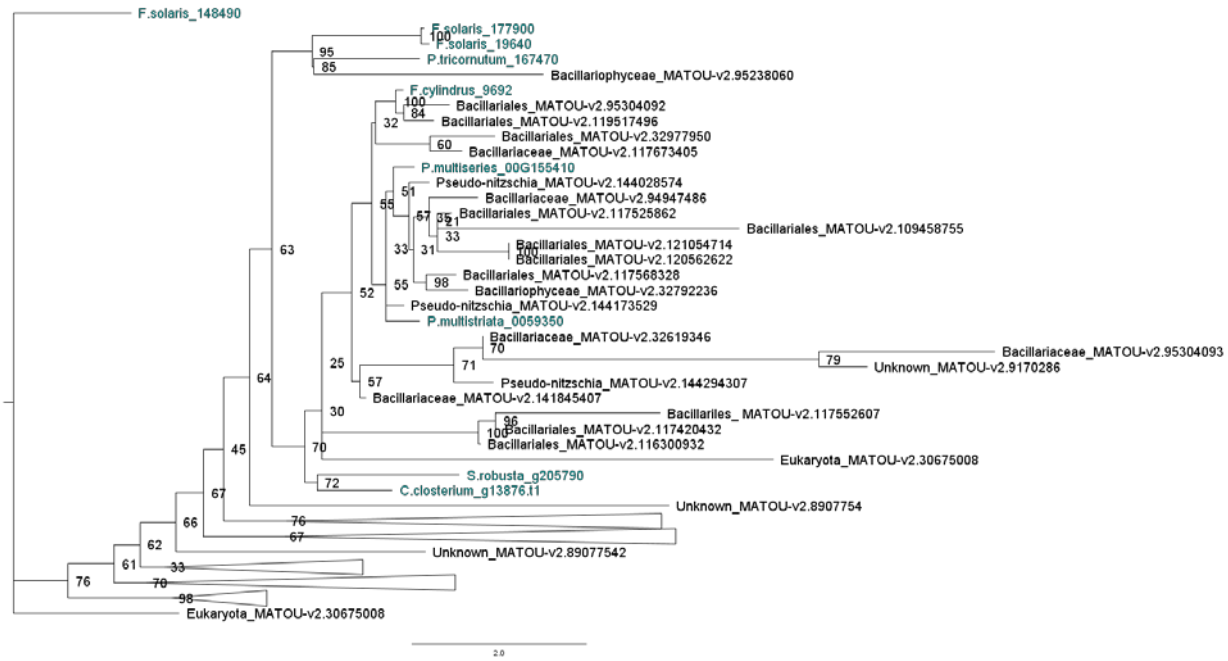

Figure 5.3. Maximum Likelihood (ML) tree of M2 genes.

Numbers at the base of each node refer to bootstrap support after 1000 replicates. Colours refer to the origin of sequences: in black sequences retrieved from the MATOU-v2 database, in pine green diatom sequences. The triangles indicate collapsed branches.

M3

The HMM search of M3 genes resulted in 170 MATOU-v2 hits, all used for the phylogenetic analysis (Table 5.2).

The phylogenetic tree (Fig. 5.4.) showed a first clade (bootstrap support: 47) with a strongly supported clade holding the *F. solaris* query and a MATOU-v2 hit annotated as belonging to the Naviculales order, together with a clade of two almost identical MATOU-v2 sequences, one ascribed to the Thalassiosirales order, and the other annotated only at the phylum level (Bacillariophyta). A low bootstrap support (value: 38) was at the base of a node that separated the first clade by a second and bigger diatom cluster; within this group, a highly supported clade (also showing high bootstrap values in the internal nodes) contained several MATOU-v2 hits that clustered together with both *F. cylindrus* and *P. multistriata* queries, while other MATOU-v2 genes fell in another highly supported clade that contained the *S. marinoi* query. Several MATOU-v2 sequences did not cluster with any of the used queries, being thus represented by collapsed branches in Fig. 5.4. However, other two distinct clades contained other query sequences. The former grouped known sequences of *P. multiseries*, *S. robusta*, *P. tricornutum*, *S. acus* and *C. closterium*. Within this big clade, overall highly supported, 10 MATOU-v2 hits were found, with annotation spanning from phylum level (Bacillariophyta) to class (e.g. Bacillariophyceae, Coscinodiscophyceae), order (e.g. Bacillariales, Fragilariales) and family (e.g. Bacillariaceae, Fragilariophyceae). Moreover, one MATOU-v2 hit, annotated as unknown *Fragilariopsis*, lied in the same cluster as the known *P. multiseries* sequence. Notably, this MATOU-v2 *Fragilariopsis*-annotated unigene did not cluster with the *F. cylindrus* sequence we used as query. This apparent ambiguity can be explained by the known genetic similarity between the two genera (*Lundholm et al., 2002*) and their phylogenetic relatedness (*Lim et al., 2016, 2018*). Finally, a clade containing known sequences of *T. oceanica* and *S. marinoi* held several MATOU-v2 sequences, possibly all belonging to centric diatoms.

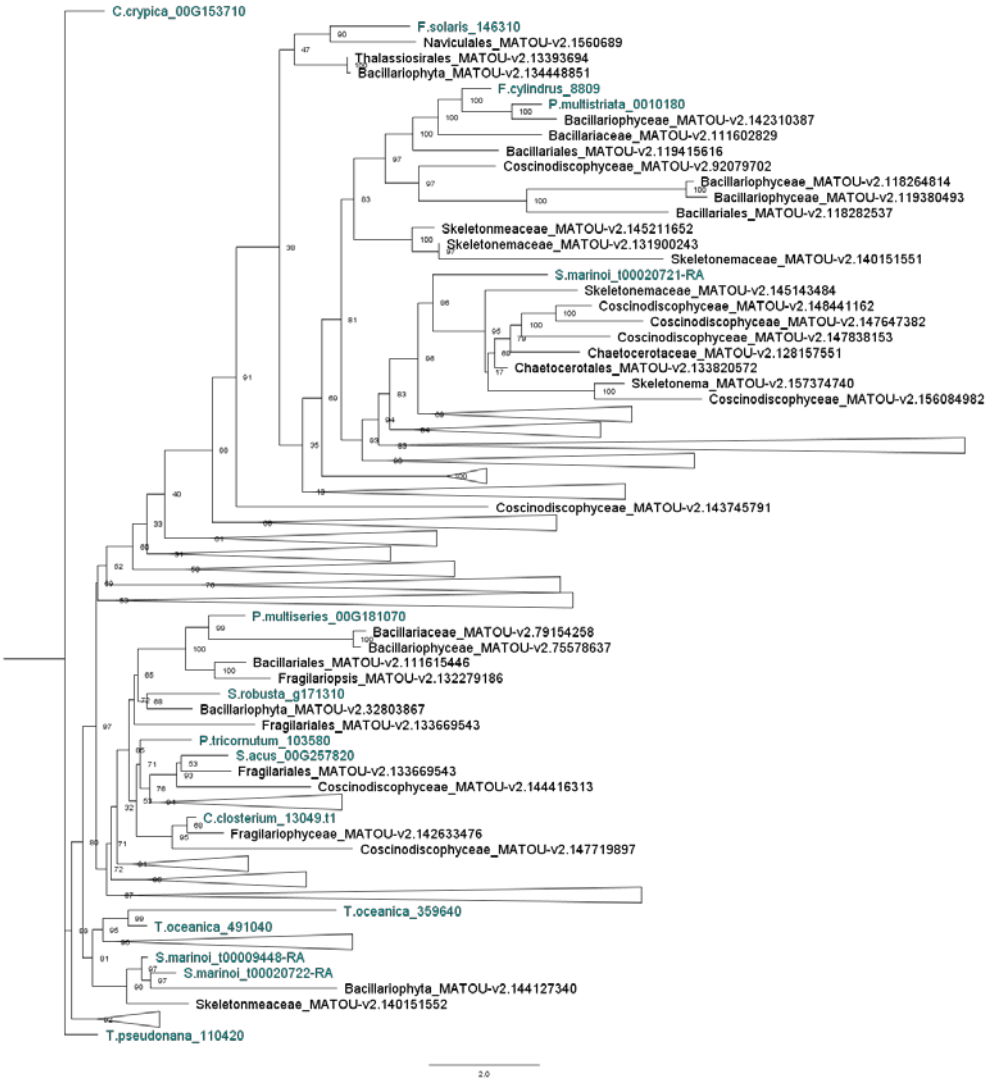

Figure 5.4. Maximum Likelihood (ML) tree of M3 genes.

Numbers at the base of each node refer to bootstrap support after 1000 replicates. Colours refer to the origin of sequences: in black sequences retrieved from the MATOU-v2 database, in pine green diatom sequences. The triangles indicate collapsed branches.

M4

The HMM search of M4 genes resulted in 41 MATOU-v2 hits, all used for the phylogenetic analysis (Table 5.2).

The tree in Fig. 5.5 shows one large clade that included both *Pseudo-nitzschia* query sequences. In particular, two MATOU-v2 hits annotated as *Pseudo-nitzschia* spp. fell next to the *P. multiseries* query. This large clade overall contained 14 MATOU-v2 sequences, mostly annotated at the order (Bacillariales) or family (Bacillariaceae) level. A second more heterogeneous and well supported clade (bootstrap value:

73) contained the two *S. robusta* queries, together with a large group of MATOU-v2 sequences that did not cluster with any query sequence. Separated by this clade there was a group containing the *C. closterium* query and three MATOU-v2 hits, one annotated at order (Bacillariales) and two at class level (Bacillariophyceae).

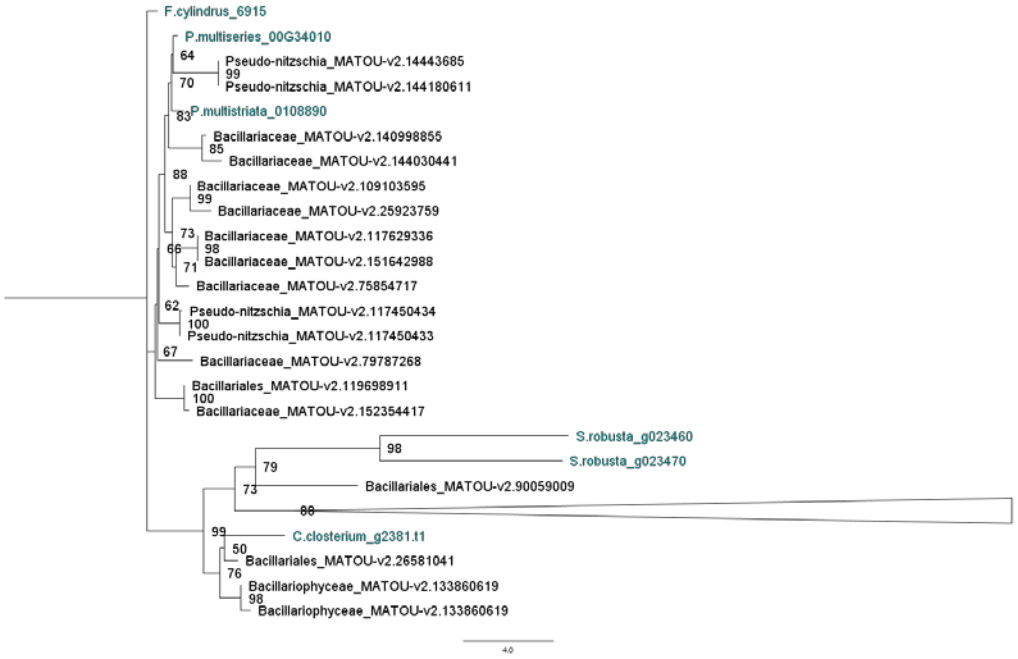

Figure 5.5. Maximum Likelihood (ML) tree of M4 genes.

Numbers at the base of each node refer to bootstrap support after 1000 replicates. Colours refer to the origin of sequences: in black sequences retrieved from the MATOU-v2 database, in pine green diatom sequences. The triangle indicates collapsed branches.

M5

The HMM search of M5 genes resulted in 312 MATOU-v2 hits, all used for the phylogenetic analysis (Table 5.2).

M5 was the only gene family for which both *P. multiseries* and *F. cylindrus* hold two copies. Interestingly, the phylogenetic tree (Figure 5.6.) shows the two copies falling in two separate clades for both species. In particular, *F. cylindrus*_6047 and *P. multiseries*_00G169490 clustered together in a strongly supported clade (bootstrap value: 99) that held several MATOU-v2 sequences, the majority of which was assigned to the Bacillariaceae family, with the exception of four hits annotated as *Fragilariopsis* spp. The other two copies of known *F. cylindrus* and *P. multiseries*

sequences lied in a separate clade (bootstrap value: 69). In particular, *F. cylindrus*_12812 clustered in a strongly supported clade (bootstrap value: 93) with two MATOU-v2 hits annotated as *Fragilariopsis* sp., as well as with two sequences belonging to Bacillariaceae. *P. multiseries*_00G12700 clustered with three MATOU-v2 *Pseudo-nitzschia* species, out of which two were annotated at the species level, representing *P. fraudulenta* and *P. arenysensis*. The clade that overall contained both *F. cylindrus_12812* and *P. multiseries_00G12700* also held the *P. multistriata* query sequence, as well as several MATOU-v2 hits annotated at order or family level (Bacillariales and Bacillariaceae, respectively). *T. oceanica* query lied in a strongly supported clade with two MATOU-v2 hits annotated as Thalassiosiraceae, while *T. pseudonana* did not cluster with other sequences. Finally, the *S. marinoi* query was found in a highly supported clade (bootstrap value: 96) that contained two MATOU-v2 hits annotated as *Skeletonema* sp., together with one Skeletonemaceae and two Thalassiosirales sequences.

Figure 5.6. Maximum Likelihood (ML) tree of M5 genes.

Numbers at the base of each node refer to bootstrap support after 1000 replicates. Colours refer to the origin of sequences: in black sequences retrieved from the MATOU-v2 database, in pine green diatom sequences. The triangles indicate collapsed branches.

SPO11-2

The HMM search of SPO11-2 genes resulted in 983 MATOU-v2 hits, out of which 635 were used for the phylogenetic analysis (Table 5.2). The phylogenetic tree (Fig. 5.7.) showed a first well supported clade (bootstrap value: 98) holding the query sequences of the centric diatoms *T. oceanica* and *S. marinoi*. A total number of six MATOU-v2 hits fell in this group; they were all annotated at the genus level, but there was not a clear phylogenetic separation between the hits belonging to

Skeletonema and the ones ascribed to *Thalassiosira*. This first clade was robustly separated from another bigger cluster. The latter included a strongly supported clade that held *F. cylindrus* as well as *P. multiseries* and *P. multistriata* query sequences; MATOU-v2 hits falling in this cluster were mostly annotated at the order (Bacillariales) or family (Bacillariaceae) level, with the exception of one *Fragilariopsis* sp. The *C. closterium* query lied separated from the *Fragilariopsis/Pseudo-nitzschia*-group, while a highly supported clade (bootstrap value: 90), containing a large amount of MATOU-v2 hits, did not hold any known query (branches collapsed in Fig. 5.7.). Another highly supported clade contained the two *S. acus* queries together with two Thalssionemataceae and one *Thalassionema* MATOU-v2 hit. Separated from this group, a strongly supported clade (bootstrap value: 86) contained centric diatoms (Coscinodiscophyceae) and two MATOU-v2 hits annotated at the phylum level (Bacillariophyta). Finally, query sequences belonging to the pennate *F. solaris* and *P. tricornutum* fell together in a highly supported clade (bootstrap value: 100), separated by *S. robusta*.

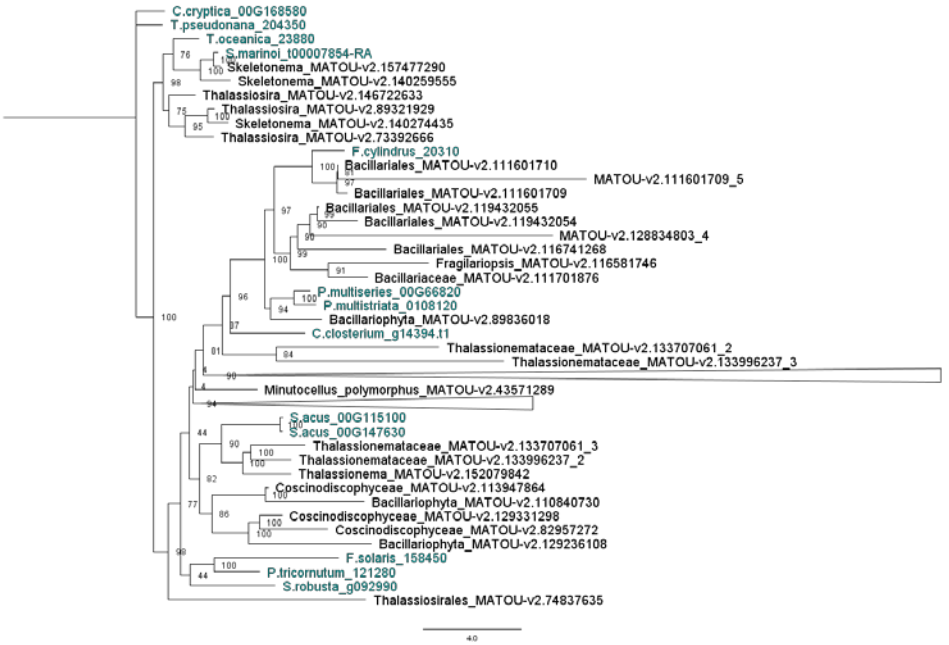

Figure 5.7. Maximum Likelihood (ML) tree of SPO11-2 genes.

Numbers at the base of each node refer to bootstrap support after 1000 replicates. Colours refer to the origin of sequences: in black sequences retrieved from the MATOU-v2 database, in pine green diatom sequences. The triangles indicate collapsed branches.

5.3.2.1 Unigene selection based on phylogenetic analysis

The visual inspection of phylogenetic trees for each marker gene family allowed to recognize which of the retrieved MATOU-v2 transcripts could be considered as the most probable homologs of the used query sequences. The selection of unigenes to be kept for the subsequent analysis was firstly performed for each marker according to the proximity of MATOU-v2 hits to the used queries: only unigenes that clustered in sufficiently supported clades (bootstrap > 50) with the known sequences were considered for the subsequent analysis. Moreover, a further selection, based on taxonomic consistency, was applied looking at phylogenetic trees altogether. In particular, as explained in section 5.2.3.2, only MATOU-v2 unigenes that clustered in strongly supported clades with the same diatom species across the phylogenetic trees of all markers were kept for the successive analysis. Notably, *Pseudo-nitzschia* and *Fragilariopsis* were the only two taxa that clustered together with MATOU-v2 homologs across all markers. Therefore, we decided to focus only on *Pseudo-nitzschia* and *Fragilariopsis* putative homologs in the subsequent analysis. Finally, an *ad hoc* consideration was made for M5, the only marker that held two copies for both *P. multiseries* and *F. cylindrus* query sequences. As described above (see section M5 in the present paragraph), the two copies clustered in separate clades for both species (Fig. 5.6). Since we did not know whether a functional difference existed between the two gene copies that belonged to the same species, we decided to select only MATOU-v2 unigenes that clustered with the *F. cylindrus* and the *P. multiseries* copies lying in the same clade as *P. multistriata*, a species for which we had experimental data showing evidence of M5 gene upregulation during sexual reproduction (see section 5.2.1.). The final number of selected genes is reported in Table 5.3.

Table 5.3. Number of selected MATOU-v2 hits after phylogenetic analysis

Marker	Number of selected hits
M1	3
M2	13
M3	7
M4	14
M5	26
SPO11-2	10

5.3.3 Expression of markers

The global distribution and the number of transcripts of each marker across *Tara* stations is shown in Fig. 5.8.

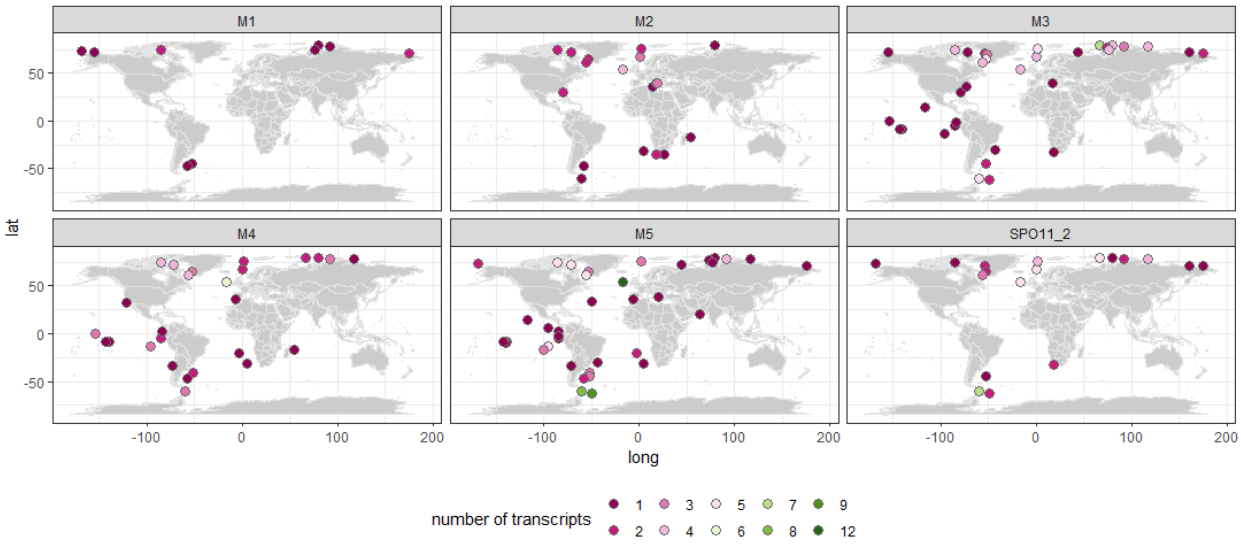

Figure 5.8. Global distribution of markers. Colours of dots indicate the number of transcripts.

All markers were found at high latitudes. Three markers were also found at lower latitudes (tropical and subtropical waters): M3, M4 and M5 were in fact found also in the tropical Pacific Ocean. Even when high latitudes were not the only regions hosting a specific marker, the Arctic and the Southern Ocean always hosted the

highest number of transcripts for each marker, suggesting that several different species were expressing the sex-induced markers. In order to detect sex events at sea, we focused on co-expression patterns. In particular, we highlighted stations in which there was the simultaneous expression of SPO11-2 and at least one of the proposed sexual markers (Fig. 5.9.A). The result consisted in 18 stations, whose distribution was constrained by the presence of SPO11-2 (Fig. 5.8); the stations extended across the whole Arctic Ocean, a region where diatoms notoriously thrive, and included one location along the North Atlantic Drift that represents the gateway from North Atlantic to Arctic Ocean (station 155). Moreover, one sample was found in the subpolar South Atlantic Ocean, and two in the Southern Ocean. Finally, one station was located at the Benguela upwelling (station 67), a region characterized by cold and nutrient rich waters.

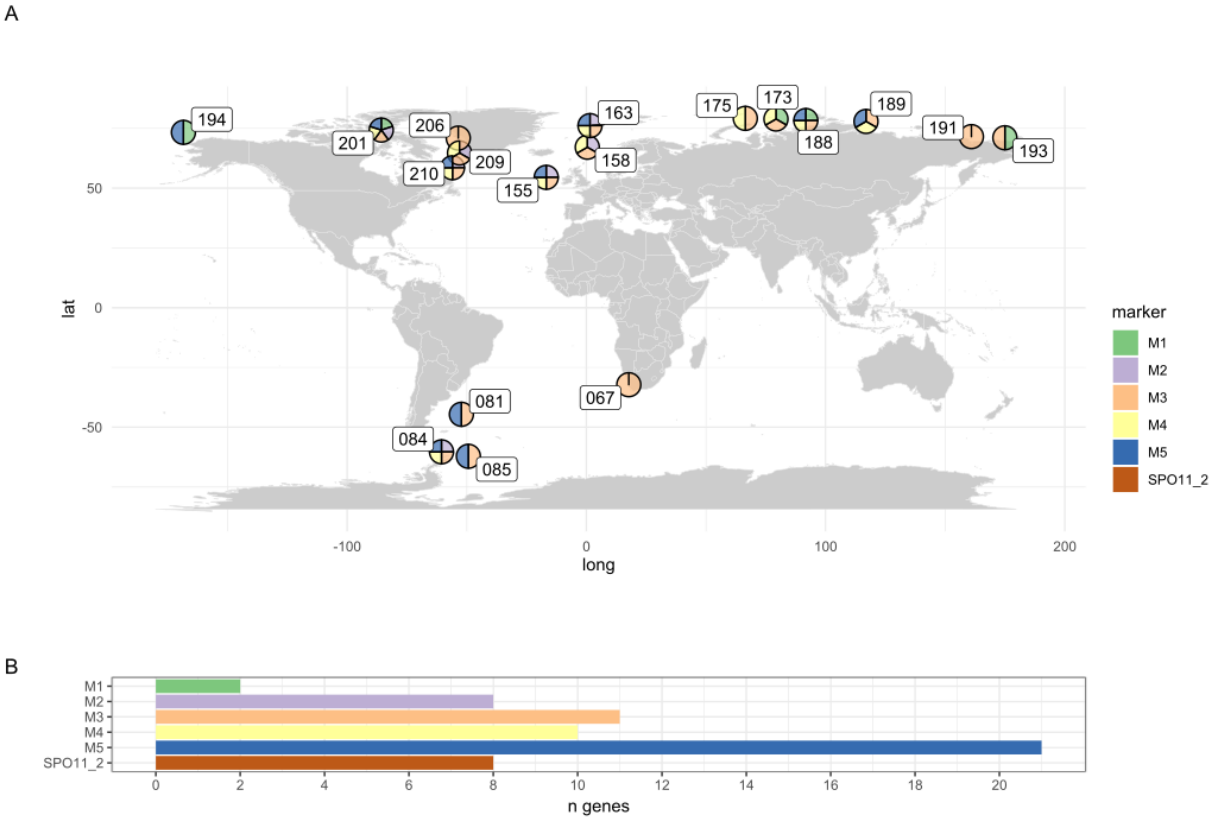

Figure 5.9. A) Geographic localization and marker composition of stations where SPO11-2 is co-expressed with at least one other marker. SPO11-2 is omitted in the pie charts as it occurs by construction in all the represented stations. B) Bar plot showing the absolute number of unigenes for each marker in the stations shown.

The number of co-expressed markers at each location varied; station 201 was the only one showing the co-occurrence of all markers, while three stations (067, 191 and 206) were characterized by the presence of only one marker, namely M3, other than SPO11-2. M3 was indeed the most widespread marker, being present in all stations except one (station 194). The rarest marker was M1, that was found only in five stations, all belonging to the Arctic Ocean, and was overall represented by two transcribed unigenes in the selected stations (Fig. 5.9.B). M2 was found in six different localities, while M4 was co-expressed with SPO11-2 in eleven stations. M5 was by far the richest marker, being represented across the selected samples by a total of 21 unigenes, and was found in ten stations. SPO11-2, occurring in all the selected stations by construction, held 8 different transcripts.

Each *Tara* station had been sampled at different depths and size fractions; we therefore explored the co-occurrence patterns at the sample level, in order to assess differences among size classes and depths. Results are shown in Fig. 5.10. and show how the co-expression of SPO11-2 and at least one other marker in one station did not imply that the same pattern was consistent across all the sampled sizes and depths. However, when a certain station was represented more than once, the same pattern of marker composition was usually found across different sizes and depths. For the subsequent analysis, we decided to investigate single samples, thus accounting for within-station differences due to distinct size and depth ranges. The bar plot also indicates that the nanoplanktonic fraction, corresponding to size fraction 3-20 or 5-20µm, was the most represented one. With one exception (station 081), it was also the only size class that showed co-expression of SPO11-2 and at least one other marker also in Deep Chlorophyll Maximum (DCM) samples. The presence of a higher signal in the nanoplanktonic size class could be an indication that the transcripts were expressed by small-sized gametes, but could also be the consequence of the fact that some cells perpendicularly passed through filters with smaller pore size during size-filtering procedure.

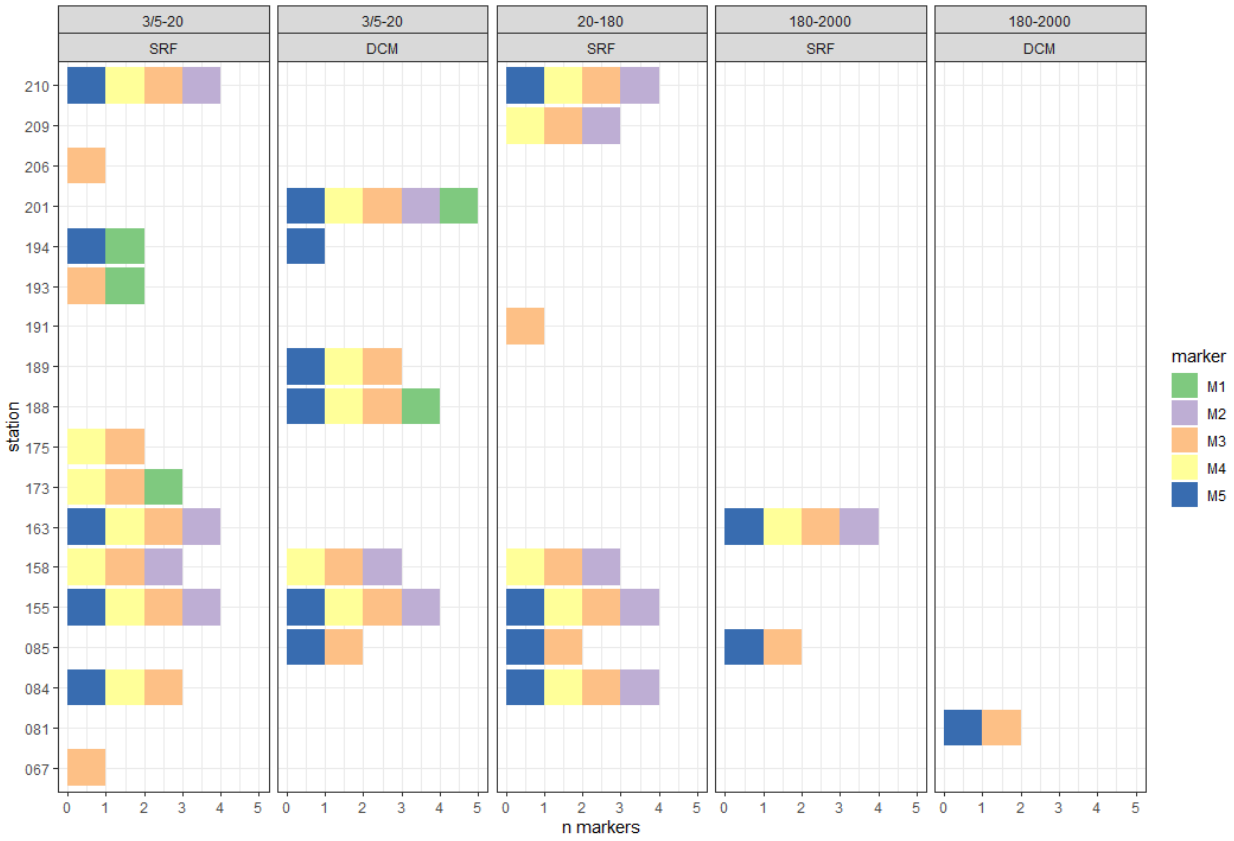

Figure 5.10. Bar plots showing the number of markers in samples where SPO11-2 is co-expressed with at least one other marker.

5.3.4 Metatranscriptomic and metabarcoding richness

Figures 5.9. and 5.10. show the composition of stations and samples in terms of marker genes co-expressed with SPO11-2. As a consequence of the filtering pipeline performed, we suspected that these markers were all putative homologs occurring in *Pseudo-nitzschia* or *Fragilariopsis* spp. The absolute number of these different homologs in each target sample was thus expected to co-vary with the number of different *Pseudo-nitzschia* and *Fragilariopsis* species expressing the detected markers at each sample. Indeed, it seemed to be the case when comparing these values, referred to as the metatranscriptomic richness, and calculated as the number of different homologs of the M1-M5 and SPO11-2 genes per sample, with the number of different OTUs found in the same samples through the analysis of metabarcoding data targeting the V4 and V9 regions of the 18S rDNA (Fig. 5.11). In particular, the bar plots show the absolute metatranscriptomic

richness, i.e., the number of distinct putative *Pseudo-nitzschia* or *Fragilariopsis* unigenes in a sample, and metabarcoding richness, expressed as the number of different *Pseudo-nitzschia* and *Fragilariopsis* OTUs (Fig. 5.11A). The correlation between these values, shown in the scatterplot in Fig. 5.11C, confirmed that the two variables were moderately correlated. The same metatranscriptomic richness was also compared with metabarcoding richness calculated over the amount of all diatom OTUs, thus including also genera other than *Pseudo-nitzschia* and *Fragilariopsis* (Fig. 5.11B). We found that the pattern of richness agreed between metatranscriptomic and metabarcoding data when looking at *Pseudo-nitzschia* - *Fragilariopsis* metaB, while the correspondence was much weaker when metabarcoding richness was calculated over all diatoms, as also indicated by the corresponding scatterplot (Fig. 5.11D), with no significant correlation. This result is a strong indication that the transcripts we selected through phylogenetic analysis were actually expressed by species belonging to *Pseudo-nitzschia* and *Fragilariopsis*. This analysis also allowed to detect samples, like the ones belonging to station 155, in which the metatranscriptomic and metabarcoding richness showed a similar pattern both when the latter was calculated over the two target genera and when done for all diatoms (especially for the V4). This result suggests that the high richness detected with metatranscriptomic data could be not a real indication of a higher gene expression of these two genera, but rather this sampling station hosted a higher richness of the whole diatom community, as observed for iron metabolism genes in Chapter III.

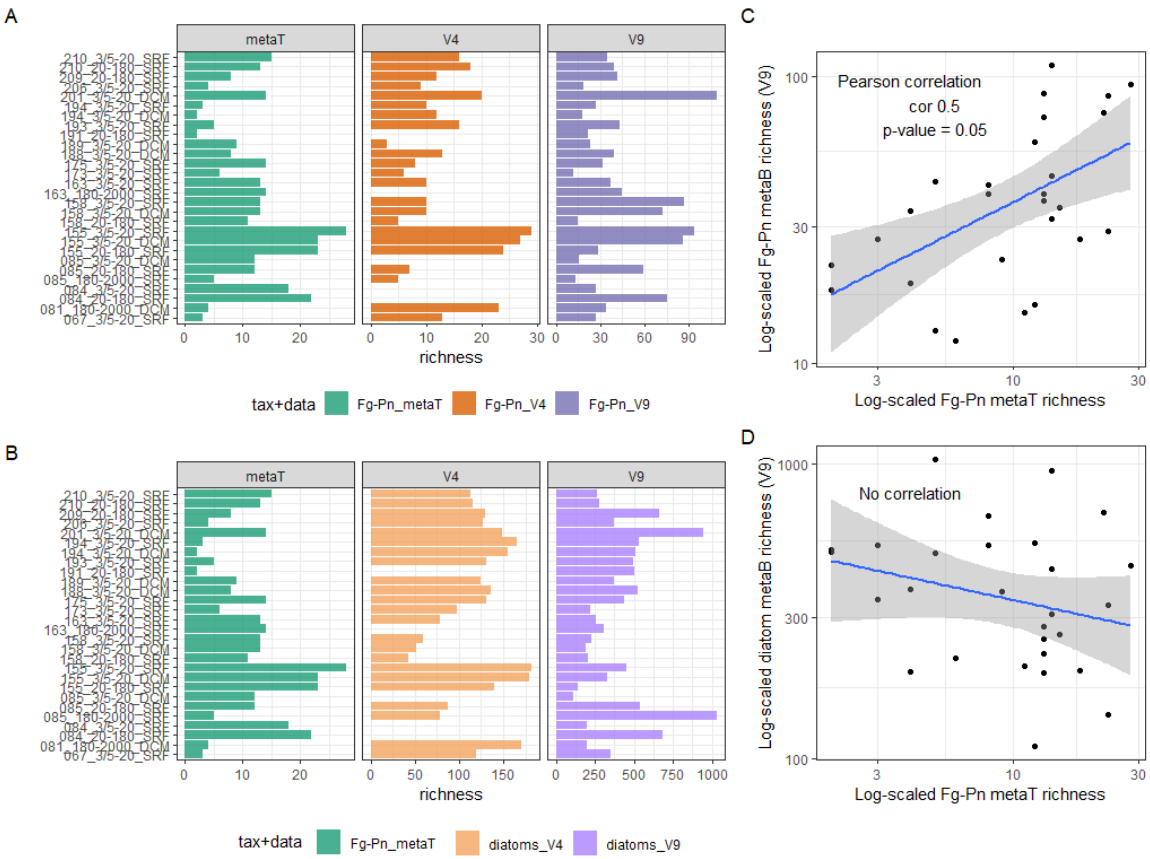

Figure 5.11. Comparison of richness patterns calculated with metatranscriptomic (metaT; green) and metabarcoding data of the V4 (orange) and V9 (violet) regions of the 18S rDNA. In A) and C) metabarcoding richness is calculated as the number of OTUs taxonomically assigned to Pseudo-nitzschia and Fragilariopsis, while in B) and D) metabarcoding richness is expressed as the number of distinct diatom OTUs in each sample. Samples are indicated in the bar plots (A, B) as the number of the station, the size fraction and the depth (SRF: surface; DCM: Deep Chlorophyll Maximum) and are represented as dots in the scatterplots (C, D).

5.3.5 Integrating metagenomic information and OTUs relative abundance

High cell density and temporal synchronization are thought to be essential to sustain planktonic populations that undergo sexual reproduction in natural environments (*Weinkauf et al., 2022*). Experiments performed on *P. multistriata* showed how individuals underwent sexual reproduction only when reaching a high cell density that favoured the spatial proximity of cells, a necessary condition for a successful chemical communication between the two opposite mating types (*Scalco et al., 2014*; *Basu et al., 2017*). No experimental evidence of this phenomenon has been provided for *Fragilariopsis* species, but the strong similarity of the general

patterns of sexual reproduction between the two genera (*Chepurnov and Mann, 2004*) suggests that a similar chemical signalling is occurring also for *Fragilariopsis* spp., as it is a crucial step in the recognition between parental cells. However, measuring the cell concentration threshold that promotes sexual reproduction in natural populations is not straightforward, as the local variation in cell density results from complex oceanic dynamics (*Borgnino et al., 2019*). Yet, we can hypothesize that if sexual reproduction is occurring at a certain location in the ocean by *Pseudo-nitzschia* or *Fragilariopsis* spp., the relative abundance of the species belonging to these two genera should be high.

Fig. 5.12 shows that three stations (155, 158 and 163), where SPO11-2 is co-expressed with at least other three marker gene families, also hosted a remarkably high relative abundance of *Pseudo-nitzschia* and/or *Fragilariopsis* OTUs. The relative contribution of *Fragilariopsis* OTUs to the total diatoms occurring in these stations was higher than 90% in station 158 sampled at both depths in the nanoplanktonic size class, while *Pseudo-nitzschia* spp. relative abundance was always greater than 60% in surface samples of stations 155 and 163, respectively in the microplanktonic (20-180 µm) and mesoplanktonic (180-2000 µm) size classes. These three stations all showed the co-expression of M2, M3 and M4, while M5 was absent in station 158. M1 occurred only in metagenomic samples in stations 155 and 163. The absence of this marker in the corresponding metatranscriptomic data for these two stations indicated that this is not a good candidate marker for sexual reproduction, as it was not actively expressed.

Figure 5.12. A) Number of sexual reproduction markers occurring in metagenomic (left) and metatranscriptomic (right) samples where SPO11-2 is co-expressed with at least one other marker. B) Abundance of Fragilariopsis (green) and Pseudo-nitzschia (pink) OTUs relatively to the total diatom abundance in the same samples, according to V4 (left) and V9 (right) metabarcoding data. C) geographical localization of stations.

Focusing on the six samples where the co-expression of marker genes by *Pseudo-nitzschia* and *Fragilariopsis* is matched by a dominance of these two genera on the whole diatom community, we examined the single transcript composition, with the aim of increasing the resolution of the investigation. The Venn diagram in Fig. 5.13 shows that the three stations overall shared 9 transcripts. A comparable number of genes were shared between stations 155 and 158 (12), 155 and 163 (10) and 158 and 163 (11). Station 158, that was geographically intermediate between the other two stations, did not hold any gene unique to this station. In contrast, more than half of the quantity of genes occurring in station 155 were uniquely expressed in this station.

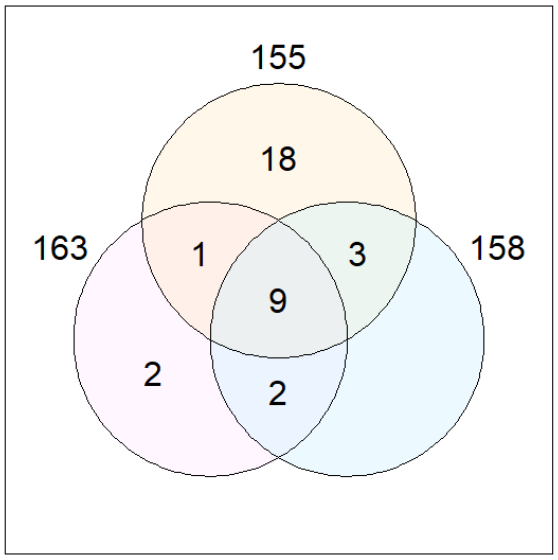

Figure 5.13. Venn diagram showing the number of transcripts shared by stations 155, 158 and 163.

When moving at the sample level (Fig. 5.14), it was clear that the high number of unigenes unique to station 155 could not only be explained by the high representativeness of this station: 10 out of the 18 transcripts uniquely expressed in station 155 would in fact show up in the Venn diagram even if we considered only one of the three samples belonging to this station, as shown by the Upset plot.

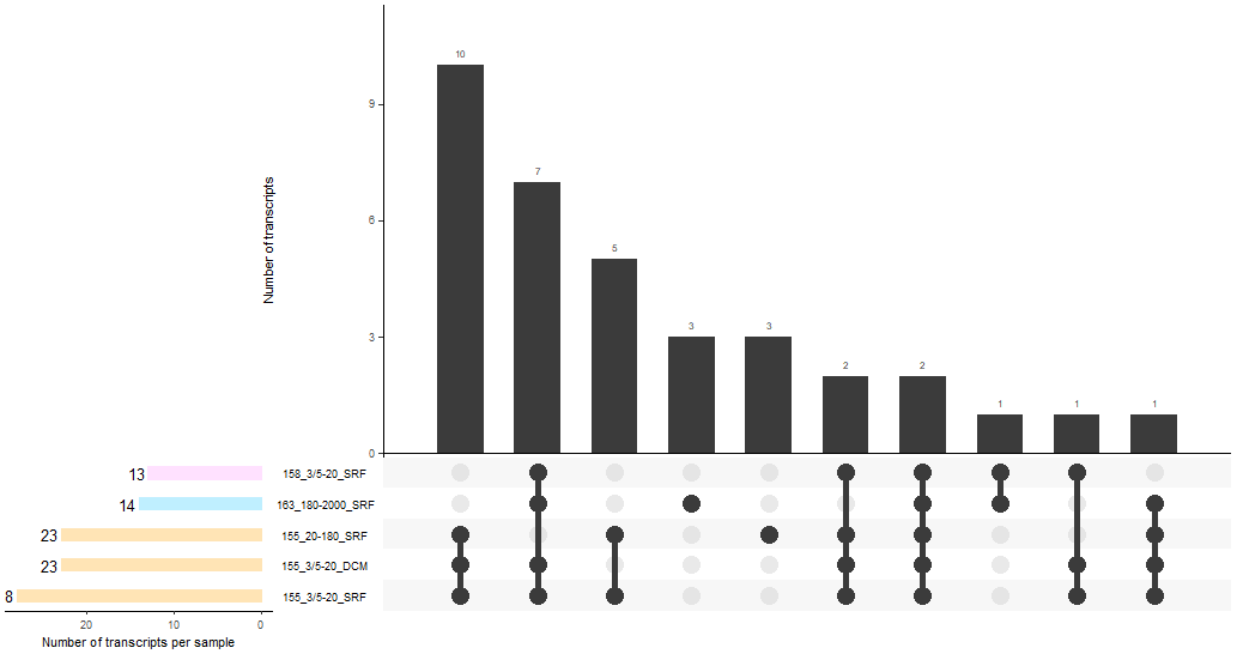

Figure 5.14. Upset plot showing the number of transcripts shared by stations 155, 158 and 163 at the sample level, i.e., accounting for different size classes and depths.

By plotting the relative abundance of the 35 unigenes occurring in the selected samples (Fig. 5.15), we obtained a glimpse on the relative expression of each transcript along the three stations identified. However, abundance of genes in the metatranscriptomic database, as retrieved from the Ocean Gene Atlas (OGA; see section 5.2.3.1.), was scaled over the total abundance of metatranscriptomic reads collected and sequenced in each sample from the whole eukaryotic community. Therefore, bar plots in Fig. 5.14 only provide a broad indication of the relative expression of one transcript respect to the others in a sample. A further effort is required to scale transcript abundances over the total abundance of diatom transcripts, or even of transcripts annotated as belonging to *Fragilariopsis* or *Pseudo-nitzschia* species, in order to understand whether the expression of unigenes of the selected sexual reproduction markers is accompanied by a lower expression of genes related to other functions. This would constitute a first contribution from natural communities of what is already been experimentally shown in *P. multistriata*, i.e., that other functions are switched off in favour of sexual

reproduction (*Basu et al., 2017*; *Annunziata et al., 2022*). The same reasoning can be applied to metagenomic data (not shown).

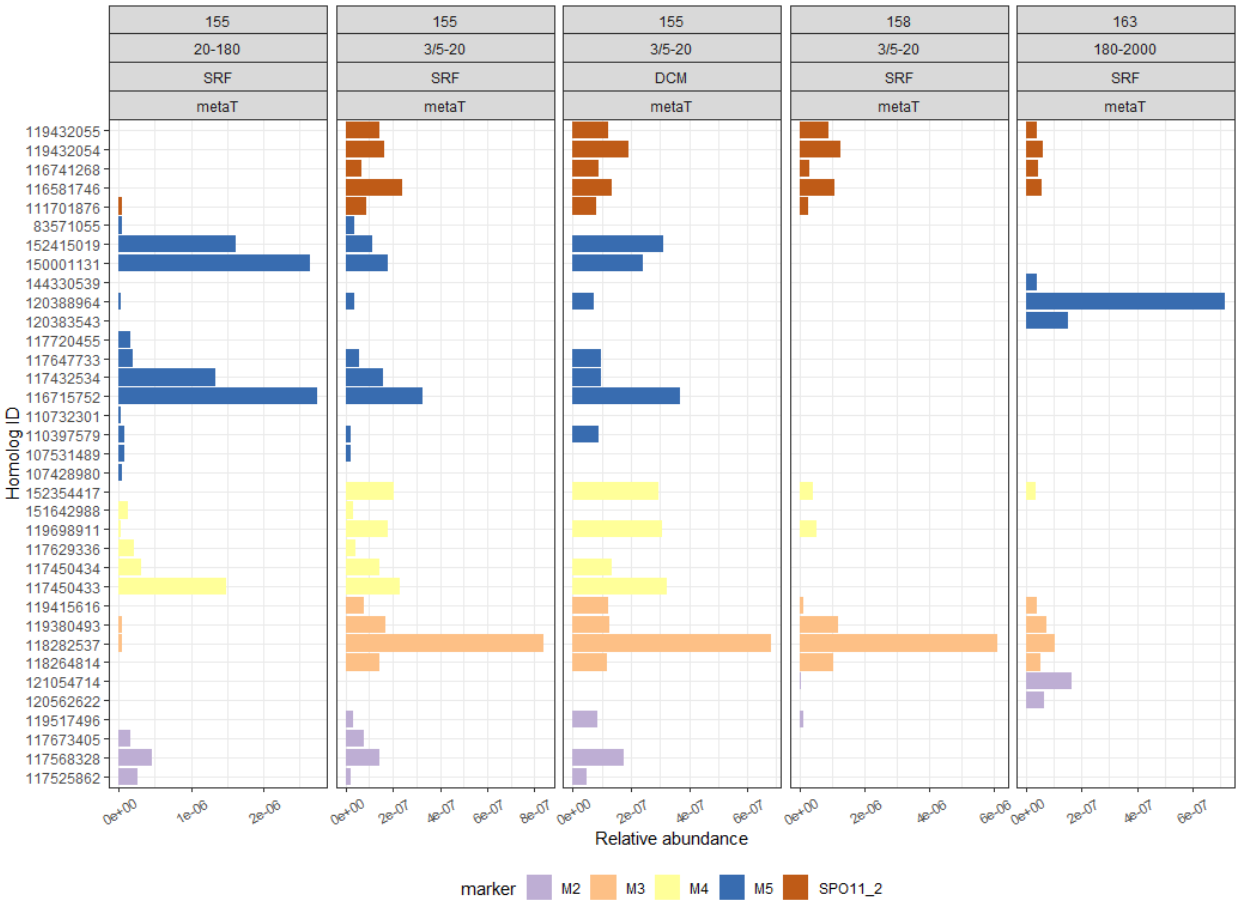

Figure 5.15. Bar plots showing the relative abundance of transcripts in samples characterized by co-expression of markers and where Fragilariopsis and Pseudo-nitzschia are the most abundant diatom genera.

5.3.6 Bloom signal

High cell density in natural environments is reached during blooms, when a remarkable acceleration of cell proliferation rate occurs. In particular, the highest values are reached at the end of the exponential growth phase of the bloom, when the sexual reproduction in *P. multistriata* has been suggested to have place (*D'Alelio et al., 2010*). Blooms and sexual reproduction are thus strongly linked events, and the evolution of bloom-forming diatoms itself has been recently proposed as a process related to sexual reproduction (*Behrenfeld et al., 2021c*). Bloom events are

overall characterized by a high production of biomass, and a rough indicator of the bulk phytoplankton biomass is chlorophyll *a* concentration. Therefore, we looked at the concentration of this pigment across *Tara* stations. The result, shown in Fig. 5.16, shows that the three stations we selected according to metatranscriptomic and metabarcoding data were characterized by a high level of chlorophyll *a* concentration compared to the average global scale value, with a higher peak reached by station 163. However, chlorophyll *a* is a proxy for biomass at the whole phytoplankton community level, and it is not possible to determine whether the concentration of this pigment is a signal of diatom biomass or is due to other species proliferation. An assessment of diatom abundance relatively to the whole phytoplankton abundance could help solving this issue.

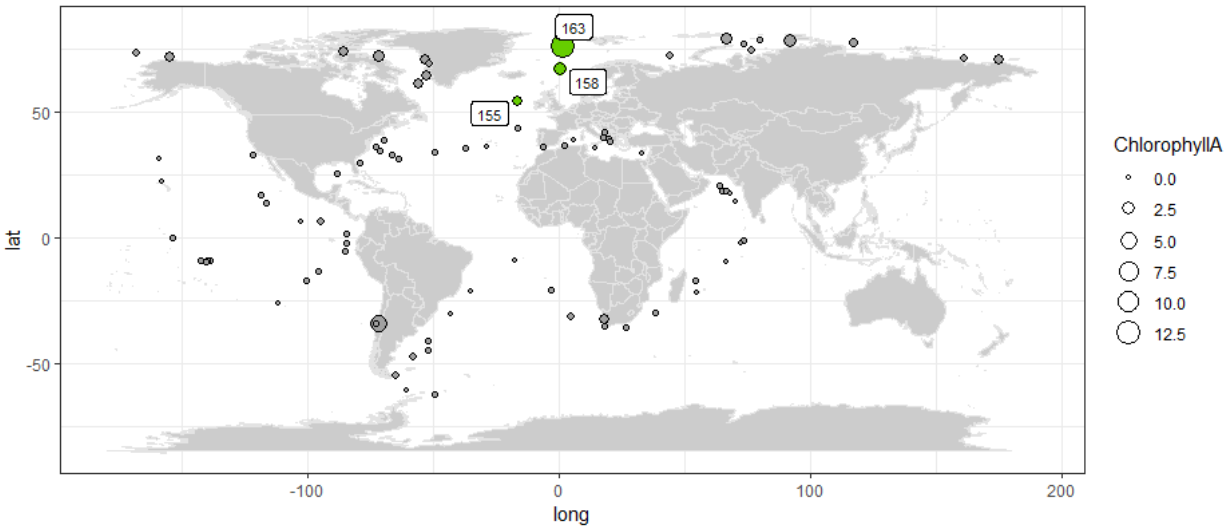

*Figure 5.16. Chlorophyll a concentration (**mg/m²**) across Tara Oceans and Tara Oceans Polar Circle stations. Size of dots is proportional to chlorophyll a concentration. Green labelled dots indicate the stations characterized by the co-expression of sexual markers and where Fragilariopsis and Pseudo-nitzschia are the most abundant diatom genera.*

5.4 Conclusion

The work presented in this chapter represented a double challenge. On the one hand, we started from the experimentally selected gene family markers to identify and detect sexual reproduction events in diatom natural populations at global

scale. While doing that, we also tested the goodness of the selected markers in order to understand if they were all equally informative. Sexual reproduction is hard to observe in the field, as it occurs during short periods, takes only a few days to complete and occurs only every few years, (*Kim et al., 2020*, and references therein). In heterothallic pennate diatoms, like *Fragilariopsis* and *Pseudo-nitzschia* species, it occurs only when different conditions are simultaneously verified. First, the population is healthy, and experiencing favourable environmental conditions. Second, cells are become too small due to recurrence of mitotic divisions, and the average cell size has to be restored. Third, cells are enough physically close to allow the exchange of chemical signals between opposite mating types. However, even when the above-mentioned conditions are met, only a small number of individuals in the population actually undergo sexual reproduction (*D'Alelio et al., 2010*). It is thus clear that assessing the occurrence of sex in natural populations is not straightforward. However, the recent advancements of genomic and transcriptomic approaches, combined with the extensive meta-omic data sampled during *Tara* Oceans campaigns, offered a unique opportunity to investigate this elusive phenomenon. We tested five different gene families proposed as markers of sexual reproduction in diatoms. Starting from the assumption that a good molecular marker is the one that allows to identify the occurrence of a process in an unambiguous and unique manner, what can we conclude about the proposed M1-M5 gene families?

To answer this question, we used the information provided by the known and functionally annotated meiotic-specific marker SPO11-2, whose occurrence at a certain location was used as an indication of the presence of samples to focus on. M1 was the weakest marker, holding only a few numbers of transcripts and being overall poorly represented (Fig. 5.8.). Moreover, several samples showed the presence of this gene in metagenomic but not in metatranscriptomic data (Fig. 5.11.). This pattern suggests that species holding this gene in their genomes do not actively transcribe it into mRNA, while a good marker should be characterized by

high expression levels during sexual reproduction. The occurrence in metagenomic and not in metatranscriptomic data is also verified for other markers, but overall in a small amount (Fig. 5.11.). The only marker whose presence at the DNA level was always accompanied by a corresponding occurrence at the mRNA level is M3. Being also diatom-specific (Fig. 5.1.) and representing the most widespread marker, as it is found in more than 90% of samples where SPO11-2 co-occurs with at least another marker (Fig. 5.11.), M3 overall emerges as the best marker family tested. M3 corresponds to the already proposed marker SIG7 (Sex Induced Gene 7; *Ferrante et al., 2019*), and contains a homologous-pairing protein 2 (hop2) domain, involved in chromosome pairing and meiotic recombination (see section 5.2.2). M3 is also the unique marker found alone with SPO11-2 in several samples. Genes belonging to the tested marker gene families should be ideally all co-expressed in a certain time from a species that is undergoing sexual reproduction. However, with the exception of one sample (Station 201), we never detected co-expression of all markers (Fig. 5.8.), also when excluding M1 from this consideration.

The integration of metabarcoding data helped to move forward in the direction of a higher resolution. Three stations, geographically close, showed interesting co-expression patterns in concomitance with a great abundance of *Pseudo-nitzschia* and *Fragilariopsis* species, the latter being a pattern not common across *Tara* samples (data not shown). The integration of environmental metadata regarding chlorophyll *a* further helped us to detect a probable ongoing diatom bloom at these localities, located along the North Atlantic Drift and including the Fram Strait, thus at regions involved in water mass exchange between the North Atlantic and the Arctic Ocean. In particular, station 155 has been sampled in May and is localized in the sub-polar North Atlantic, a region that hosts the largest annual phytoplankton bloom in the global ocean (*Behrenfeld et al., 2019*).

To conclude, what we retrieved from meta-omic, metabarcoding and chlorophyll *a* alone are fragmented and imperfect information. However, their integration

suggests that in the three selected stations sexual reproduction is occurring by *Fragilariopsis* and *Pseudo-nitzschia* species. This result is in line with the known rarity of the sexual reproduction phenomenon. It would be particularly interesting to provide an estimate on how rare sex is in pennate diatoms at sea. Although we could not make this estimation, the pipeline presented here lays the foundation for addressing this question in future experiments. It would be too hazardous to claim that we validated the panel of sexual markers for all diatoms. The required stringent filtering pipeline, built with the aim of finding a compromise between the loss of data and a robust taxonomical assignation of genes, eventually brought us to focus on only two genera, i.e., *Pseudo-nitzschia* and *Fragilariopsis*. The three selected stations all hold SPO11-2 in co-expression with markers M2, M3 and M4 (M5 is absent in station 158); we could thus indicate these three gene families as putative sexual reproduction makers for pennate diatoms.

The integration of more experimental data, as well as the inclusion of genes from transcriptomic databases, like the Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP), could increase the resolution of phylogenetic trees and allow a better evaluation of the taxonomic nature of MATOU-v2 unigenes. A worth mentioning aspect is, in fact, that in most cases we did not find high resolution taxonomic annotation of MATOU-v2 genes. However, *Tara* Oceans and *Tara* Oceans Polar Circle omic data are not limited to the metagenomic and metatranscriptomic information we explored in the present chapter. A recent effort, in particular, was carried out to reconstruct and manually curate metagenome-assembled genomes (MAGs) from eukaryotic metagenomic reads from 143 *Tara* Oceans stations (*Delmont et al., 2022*). We considered all the genes studied here as single-copy genes; we actually do not know that. Mapping the sequences of the retrieved MATOU-v2 unigenes on *Pseudo-nitzschia* and *Fragilariopsis* MAGs could help us to better understand the nature of the genes, as well as better resolve their taxonomy.

As mentioned in section 5.3.4., we could not make the most of the relative abundance data from metatranscriptomics because we need to scale the obtained values on the total abundance of diatoms. Moreover, a further step could be to compare the ratio between functions in the selected stations, in order to understand if metabolic-related genes of *Pseudo-nitzschia* and *Fragilariopsis*, like the ones involved in processes of uptake and transport of nutrients, as well as photosynthesis-related genes, are less expressed in stations where individuals are undergoing sexual reproduction, as experimentally demonstrated through RNA-sequencing studies (*Basu et al., 2017*; *Annunziata et al., 2022*). Furthermore, besides the molecular information, *Tara* Oceans expeditions collected and produced a large amount of morphological data. Although the microscopy identification of diatom sexual stages is not easy, observations of the morphology of collected samples could provide a further indication of whether sexual reproduction is occurring at a given location.

In conclusion, sexual reproduction is a pivotal process in life cycle of diatoms and has fundamental implications for population dynamics and community structure of these unicellular eukaryotes, although further steps are required to fully understand the mechanisms involved in this complex process. The study presented here added a contribution on the detection of sexual events at sea in pennate diatoms, with a special focus on *Fragilariopsis* and *Pseudo-nitzschia* species.

Chapter VI

Thesis Summary and Future Perspectives

6.1 Thesis scope and main results

The general scope of my thesis was to explore diatom ecology through the unprecedented amount of data generated through the *Tara* Oceans expeditions. During this investigation, I encountered both methodological and conceptual challenges. Environmental omics is a promising field and the number of studies exploiting omic data is exponentially increasing, with several implications. On one hand, the scientific community has the urgent need of rapidly developing and assessing standardized procedures to deal with this type of data, from the collection to the wet lab processing and the bioinformatics pipelines. On the other hand, the advent of this new large amount of data offers extraordinary opportunities to revamp long-standing scientific debates with new approaches and to formulate new biological questions. In this context, in my PhD thesis I explored concepts of diatom taxonomical and functional diversity, community structure, biogeography and ecology, by integrating different approaches and methodologies, as well as different types of data.

Starting from a macroecological perspective (Chapter II), I showed how inherent properties seem to rule the main structural patterns across diatom communities and I highlighted the deviation from the general observations obtained with functional data (i.e., metatranscriptomics). The results propose the integration of metatranscriptomic information in macroecological studies as they add a new perspective on the universal laws that rule large scale patterns in ecology. Diatom communities were also characterized through traditional alpha and beta diversity

indices and taxonomical composition. This approach allowed the description of diatom communities across an oceanic transect of major interest in the context of climate change (Chapter III). I demonstrated how diatoms inhabiting the North Atlantic and the Arctic Ocean waters represent two separate communities yet hosting intra-community heterogeneity, and showed the existence of a subpolar zone of transition between the two oceans. Biogeography and biodiversity patterns were also studied at a larger spatial scale, exploiting the entire spatial coverage of *Tara* Oceans expeditions and including samples from other projects, and at a higher taxonomical resolution, focusing on the diatom genus *Pseudo-nitzschia* (Chapter IV). In particular, I described here a methodology that starts from reference sequences to exploit metabarcoding data overcoming the limitations of the use of ribotypes processed in the form of OTUs or ASVs; in this context, I showed how the integration of curated reference data sets with metabarcoding information can provide insight on the biogeography and biodiversity assessment at the genus and species level. I further exploited *Pseudo-nitzschia* ecology by integrating the information from prokaryotic OTUs (Chapter IV). In particular, I applied network analysis to study the putative interactions between the diatom and bacterial components of microbial communities sampled across the global ocean. Thanks to the size fractionated nature of *Tara* samples, I could discriminate between free-living and putative algae-attached bacteria. By constructing correlation networks and performing downstream analysis, I could characterize free-living and particle-attached prokaryotic communities, whose separation reflected in clear taxonomical differences and patterns that have now to be characterised from a functional point of view by means of laboratory experiments.

Finally, one of the main methodological advancements of my thesis was the proposition of a robust pipeline to identify genes in metatranscriptomic data (Chapter III and Chapter V). With the aim of describing general patterns of functional variation of diatom communities across oceanic regions, at the beginning of my PhD study I exploited the *Tara* Oceans unigenes by looking at their

Pfam annotation, as this information was already provided in the data set. However, the obtained results were generally insufficient, as one Pfam can be assigned to more genes and a single gene can correspond to several Pfams. Moreover, often describing a single protein domain and not the whole gene function, I found Pfam-based information resulting in an overall low degree of functional completeness. While trying to overcome these limitations, I established a strong collaboration with Dr Camilla Borgonuovo (Stazione Zoologica Anton Dohrn, Italy) that originally developed the pipeline that I presented in Chapters III and V (*Borgonuovo, 2019*); in particular, this collaboration allowed me to contribute to the final design of the pipeline and to widen the spectrum of processes which it can be applied. This pipeline consists of progressive filtering steps that, starting from a set of genes of interest, permitted the robust detection of them in large metatranscriptomic data sets. Following genes detection, the pipeline was applied to studying gene expression in the context of two different processes: iron metabolism and sexual reproduction. In the first case, I could describe the spatial variation in functional expression of genes related to iron uptake, transport and storage, whose transcriptomic activity was also profoundly shaped by the taxonomy. I found that the two main groups of diatoms, i.e., Pennates and Centrics, express different genes involved in uptake and transport of iron in their cells, and that this expression changes along the currents that bring warm water from the North Atlantic into the polar Arctic ecosystem. In the second case, the pipeline allowed the assessment of a first biogeography of diatom sexual reproduction expressed genes and the identification of three sampling stations where putative sexual events were occurring by species belonging to either *Pseudo-nitzschia* or its closely related genus *Fragilariopsis*.

6.2 Thesis summary

The present thesis is structured to cover the investigation of diatom biodiversity, biogeography and functioning starting from community level and then moving

towards lower taxonomic ranks, in particular focusing on the ecologically relevant genus *Pseudo-nitzschia*.

In the first chapter I provided a comprehensive overview of the background context for the topics addressed by this thesis. I reviewed old and new concepts of phytoplankton and diatom biodiversity and biogeography and the processes behind what we observe, as well as describing the functional role of diatoms in the context of the main biogeochemical cycles. I also provided a more in-depth description of the genus *Pseudo-nitzschia* by first highlighting its relevance from a socio-ecological point of view. I then illustrated its diversity and biogeography as well as the life cycle traits of its species. In conclusion, I described the advent of the multidisciplinary field of environmental omics, illustrating the main contribution to this discipline provided by *Tara* Oceans project and its enormous potential to advance our understanding of plankton ecology. In this context I also provided a description of the main high throughput sequencing (HTS) data I exploited along the chapters.

Chapter II provided a first description of diatom communities sampled across the global ocean. I studied diatom communities from a macroecological perspective and revealed emergent patterns in diatom communities' structuring. By first providing a general overview of the studied data set I showed that diatom communities detected through metabarcoding, metagenomic and metatranscriptomic data are overall more abundant and diverse at high latitudes and in cold waters, making them one of the main contributors of the whole sampled plankton community in polar regions. Diatom biodiversity strongly correlated with temperature and salinity values (negative correlations) and to a lesser extent with nitrate and phosphate (positive correlations). A high number of rare entities in diatom communities emerged from the analysis of species abundance distribution patterns as obtained by metabarcoding, metagenomic and metatranscriptomic data. The analysis of the relationship between the main statistical descriptors of abundance confirmed one of the main macroecological

laws, i.e., Taylor's law, and further clarified the long-tailed and right-skewed nature of the distributions. In the last part of the chapter, I related the total diatom abundance of a station to four main biodiversity indices, namely richness, evenness, dominance and rarity. I showed how a higher number of species in a location reflects into a high richness, dominance and rarity and a lower evenness. The simultaneous decrease of evenness and increase of rarity, observed for both metabarcoding and metagenomic data when N increased, indicates that a greater abundance corresponds to an increment of rare entities that overall amplify the uneven distribution between the dominant and non-dominant components of the community. This observation was not true for metatranscriptomic data, suggesting the existence of different mechanisms that rule the patterns of expressed transcripts.

Chapter III was still aimed at exploring diatom communities but on a regional rather than a global scale: I decided to focus on populations that travel with the oceanic currents in the North Atlantic Ocean and that bring warm and salty water into the polar Arctic Ocean. The aim is reconciling the classical macroecological approach based on analysing geographic patterns with the fact that populations are transported (i.e., move from a landscape to a *seascape* framework). This is even more important for understanding the Arctic Ocean response to climate change since this basin is continuously fed by water masses from the adjacent basins. I showed how diatom communities inhabiting the two oceans are different, integrating the information of two metabarcoding markers, i.e., the V4 and the V9 regions of the 18S rDNA. I also highlighted the existence of a transition zone between the two systems, where OTUs belonging to both oceans occur in a higher percentage yet presenting its own properties; this is relevant as the amplitude of this region is expected to increase in the next future and extend poleward. In this chapter I also showed how the spatial variation in diatom communities was accompanied by a variation in functional expression of genes related to iron metabolism, whose activity was also profoundly shaped by taxonomy. I explored

the main relation between genes involved in three different mechanisms of iron uptake and transport, namely the one based on general divalent transporters (NRAMP and ZIP), the reductive uptake (FRE, FET/MCO and FTR genes) and the siderophore-mediated transport (ISIP1 and FBP genes); two genes involved in the storage of this micronutrient were also included (FTN and ISIP3). Although a similar pattern was retrieved along the study site when looking at process level, I found that the two main taxonomical groups of diatoms, i.e., Pennates and Centrics, express different genes involved in iron uptake and transport, and that this expression changes along the current system that flow from the North Atlantic to the Arctic Ocean, that is, I described the functional turnover occurring during the journey to the Arctic.

Chapter IV and Chapter V were addressed to the study of species belonging to the genus *Pseudo-nitzschia*. In Chapter IV I used a taxonomically curated reference database and looked at single ribotypes to explore *Pseudo-nitzschia* species biogeography and ecology at global scale. A first analysis based only on species occurrence across sampling stations in the context of several projects allowed the detection of several species showing a temperate-tropical distribution; moreover, I found that species occurring at poles were often found elsewhere, and that several *species* were ubiquitous; I also related biogeographical patterns to the main environmental parameters, addressing the degree of tolerance of each species towards the main nutrient and physical parameters across samples. A subset of these samples, corresponding to the ones collected in the context of *Tara* Oceans expeditions, were further explored using the information on relative abundance of each *Pseudo-nitzschia* species to the total genus-level community. This information enriched the biogeographical assessment and showed how species that occur across a wide spatial range are locally abundant only in specific regions of the ocean. The Arctic Ocean appeared as dominated by the toxic *P. australis* and *P. multiseries*, also found in subpolar regions of both hemispheres, and cold waters in general, together with *P. fraudulenta*; Southern Ocean stations were mainly

dominated by *P. delicatissima*, that also occur in subpolar samples, together with *P. turgidula*. The tropical Pacific Ocean was dominated by species belonging to a group of species impossible to be separately identified through the used metabarcoding marker (i.e., V4), and namely *P. americana*, *P. caciantha*, *P. cirumpora*, *P. granii* and *P. linea*. *P. brasiliana* and *P. lineola* are also highly abundant in the Pacific Ocean, together with the cosmopolitan *P. micropora*. The Mediterranean Sea is dominated by *P. mannii*, while other stations host a more even distribution of species. A further step in the study of *Pseudo-nitzschia* ecology involved the investigation of the co-variation patterns of *Pseudo-nitzschia* species with prokaryotic OTUs. A first result of this analysis is in line with the experimentally shown species-specific nature of the associations between *Pseudo-nitzschia* and bacterial OTUs. With the aim of distinguishing between bacteria inhabiting the water surrounding each algal cell and the ones that actually live attached to the surface of diatoms, I separately studied how *Pseudo-nitzschia* species co-varied with bacteria sampled in the free-living size fraction (0.22-3 µm) and in the particle-attached one (20-180 µm). The comparison between the taxonomic information relative to both the free-living and the putative algae-attached bacteria allowed to discriminate between the two different communities of bacteria and encourage the validation through experimental studies, also necessary to understand the nature of the putative interactions.

Finally, in Chapter V I explored a set of genes recently proposed as molecular markers for sexual reproduction in diatoms. I used this panel of marker genes with the scope of detecting sexual reproduction events in diatom natural population at global scale, with a special focus on *Pseudo-nitzschia* and *Fragilariopsis* species. In this chapter I applied the same functional pipeline described in Chapter III (see section 6.1) and investigated the co-occurrence patterns of transcripts belonging to the target genes along *Tara* Oceans and *Tara* Oceans Polar Circle sampling stations. The obtained results were corroborated by the integration of information from metabarcoding data in terms of OTUs abundance and richness, together with

chlorophyll *a* concentration as a proxy of phytoplankton biomass; altogether, the results suggested the identification of three *Tara* sampling stations as localities where sexual reproduction was having place by *Pseudo-nitzschia* and/or *Fragilariopsis* species, all associated to very high biomasses.

6.3 General comments and concluding remarks

Meta-omic data did not exist until a few decades ago, yet they are now pivotal for a successful understanding of population dynamics and ecological processes of the microbial world; however, the approaches and tools to analyse them are still in evolution. Although under a constant improvement, there is still no unique consensus on standardized methodologies to deal with meta-omics, from sampling protocols to experimental and bioinformatics procedures. The necessity of finding common strategies and pipelines is made even more urgent if considering the potential that these data have to provide powerful insight in the context of climate change. With respect to the present thesis, the methodological approaches I proposed represent useful tools to investigate biogeography and diversity concepts in marine plankton; however, they are not exempt from limitations.

One strong limitation in this study relies to an inherent feature of high throughput sequencing (HTS) data; in particular, it concerns the progressive subsampling that occurs while generating this type of data. The clean sequencing reads, ready to be used for downstream analysis, are in fact the result of progressive subsampling steps, that start from the collection of data in field and include the wet lab techniques of nucleic acid extraction, amplification and sequencing, followed by trimming and quality filtering bioinformatics pipelines. The obtained reads thus inevitably represent only a random subset of the real population. Furthermore, in order to compare different samples, sequencing outputs in the form of absolute read counts are usually converted to relative abundance values through different types of normalization. One of the most used is the total sum scaling (TSS), which

divides read counts by the total count in each sample; this conversion from absolute to relative numbers makes data compositional, i.e., representing proportions of some whole (*Aitchison, 1982*). However, it is not possible to know the real link between absolute abundance of a taxon in the environment and its relative abundance after HTS, and although TSS is useful at removing technical biases related to different sampling depths across samples, the increasing or decreasing of a taxon in a sample can be artificially produced through this technique, ultimately biasing downstream comparisons between samples (*Gloor et al., 2017*; *Espinoza et al., 2020*). Alternative normalization techniques have been proposed as possible solutions to this problem, like the use of log-ratio transformations (*Gloor et al., 2017*; *Espinoza et al., 2020*) or the development of a normalization based upon empirical Bayesian approaches (*Liu et al., 2020*). A more direct option involves the use of spike-in controls, i.e., nucleic acids with known absolute abundance that serve as internal standards to be introduced into samples before nucleic acid extraction and sequencing (*Satinsky et al., 2013*; *Crossette et al., 2021*). The lack of completeness of HTS-derived data sets is strictly linked to another common problem in ecology, i.e., the presence of zeros in community matrices: Does each zero mean the real absence of a species or a gene in a sample? Under the assumption of global ubiquity, fundamental for several current macroecological theories, all organisms occur in the sampled system and zero values are only result of low sampling effort. However, zeros can also represent species that are truly absent from the community because not adapted to the ecosystem. Statistical approaches have been proposed to identify the nature of the observed zeros in community matrices (e.g., *Blasco-Moreno et al., 2019*) and analytically account for their presence (e.g., *Kumar et al., 2018*; *Martino et al., 2019*).

The present study is also strongly limited by the use of only reference-based approaches to identify diatom OTUs, ribotypes and genes. The taxonomical annotation of each DNA or RNA fragment required in fact the identification of its homologs by comparison with genomic or transcriptomic references. Therefore,

the occurrence in available databases of reference genomes and transcriptomes and their associated taxonomic assignation strongly affected the resolution of the exploited meta-omic data. The enormous quantity of unknown taxa emerging from almost all systematic omics-based overviews of microbial organisms reflects the incompleteness of these databases and leave us with a high quantity of organisms and genes without a name or an associated function. In the case of diatoms, only a few genomes have been sequenced. Two of them belong to species elected as model organisms for diatoms as they are easily cultivable and rapidly growing in laboratory conditions, i.e., *Thalassiosira pseudonana* (Armbrust et al., 2004) and *Phaeodactlyum tricornutum* (Bowler et al., 2008). Another species of the genus *Thalassiosira* was then sequenced (Lommer et al. 2012), namely *T. oceanica*, an ecologically relevant diatom adapted to iron-limited regions. In the last five years an increasing effort led to the sequencing of the cold-adapted *Fragilariopsis cylindrus* (Mock et al., 2017), the sexual reproduction model organism *Pseudo-nitzschia multistriata* (Basu et al., 2017), the toxic *Pseudo-nitzschia multiseries* (Basu et al., 2017) and *Skeletonema costatum* (Ogura et al., 2018), the biofuel-producers *Fistulifera solaris* (Tanaka et al., 2015) and *Cyclotella cryptica* (Traller et al., 2016). The most recent effort added a benthic diatom to the list, i.e., the pennate *Seminavis robusta* (Osuna-Cruz et al., 2020), to reveal genome adaptation to the benthic lifestyle. In addition, 92 different diatom transcripts are included in the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP; Keeling et al., 2014). Notwithstanding the mentioned resources, the reference catalogue is still far to sufficiently cover the diversity of diatoms, that, holding a number of species ranging from 4,000 to 30,000 according to recent estimates (Guiry, 2012; Mann and Vanormelingen, 2013; Malviya et al., 2016), represent one of the most diverse unicellular microalgae. Constructing a more comprehensive genomic and transcriptomic resource for diatoms will thus substantially improve the quality of meta-omic data, which will in turn profoundly impact functional and comparative genomic studies. In parallel, reference-free methods are being developed to overcome the limitations of lack of references. Basing on the assumption that the

depth of sequencing coverage, i.e., the number of reads covering a portion of a genome, strongly co-vary among samples when reads belong to the same genome, reference-free methods cluster hundreds of loci that follow the same variation of sequencing coverage between samples and ascribe them to the same genome. This methodology enabled the incorporation of population genomic approaches with the use of metagenomic samples (*Laso-Jadart et al., 2020*) and is also the rationale behind the construction of metagenome assembled genomes (MAGs), that are key step toward the detection of new species, taxonomic profiling and downstream functional analysis (e.g., *Delmont et al., 2022*).

One of the great strengths of *Tara* Oceans expeditions is its wide spatial scale. By covering the whole global ocean, this project allowed the comparison of samples collected from distant and different habitats and environments. However, a major limitation of *Tara* Oceans is the absence of the time dimension: each sample collected represent a one-time snapshot of the planktonic community that is the result of complex processes driven by strong seasonal changes that govern the dynamics of marine systems. An ongoing step in meta-omic studies is therefore the advent of long-term genomic observatories, whose standardized data collection and processing is expected to generate high-quality biodiversity data in the long term (e.g., European Marine Omics Biodiversity Observation Network; EMO BON).

The present thesis Figures as a contribution to the emergent field of environmental omics; within this thesis I used principles, techniques, theories and notions from theoretical ecology, molecular biology and oceanography to develop and apply methodologies to extract knowledge from omic data of diatoms, one of the key components of marine phytoplankton. Overcoming the limitations described in the present chapter will certainly provide the means for a better understanding of diatom and planktonic dynamics that will in turn enable a better evaluation of their impact on biogeochemical cycles and food webs, currently facing the anthropogenic climate change. Furthermore, I believe that the ongoing effort in

establishment synergies among disciplines to deal with the unprecedented amount of data obtained by these new technologies will provide powerful insight to answer new emerging biological questions.

References

Aagaard, K., & Carmack, E. C. (1989). The role of sea ice and other fresh water in the Arctic circulation. Journal of Geophysical Research: Oceans, 94(C10), 14485-14498.

Abirami, B., Radhakrishnan, M., Kumaran, S., & Wilson, A. (2021). Impacts of global warming on marine microbial communities. Science of The Total Environment, 147905.

Achterberg, E. P., Steigenberger, S., Marsay, C. M., LeMoigne, F. A., Painter, S. C., Baker, A. R., ... & Tanhua, T. (2018). Iron biogeochemistry in the high latitude North Atlantic Ocean. Scientific reports, 8(1), 1-15.

Aguilar-Islas, A. M., Hurst, M. P., Buck, K. N., Sohst, B., Smith, G. J., Lohan, M. C., & Bruland, K. W. (2007). Micro-and macronutrients in the southeastern Bering Sea: Insight into iron-replete and iron-depleted regimes. Progress in Oceanography, 73(2), 99-126.

Alatalo, R. V. (1981). Problems in the measurement of evenness in ecology. Oikos, 199-204.

Alberti, A., Poulain, J., Engelen, S., Labadie, K., Romac, S., Ferrera, I., ... & Wincker, P. (2017). Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. Scientific data, 4(1), 1-20.

Albright, L. J., Yang, C. Z., & 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), 215-225.

Alexander, H., Jenkins, B. D., Rynearson, T. A., & Dyhrman, S. T. (2015). Metatranscriptome analyses indicate resource partitioning between diatoms in the field. Proceedings of the National Academy of Sciences, 112(17), E2182-E2190.

Allen, M. R., de Coninck, H., Dube, O. P., Hoegh-Guldberg, O., Jacob, D., Jiang, K., ... & Zickfeld, K. (2018). Technical summary. In Global warming of 1.5° C: An IPCC Special Report on the impacts of global warming of 1.5° C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty (pp. 27-46). Intergovernmental Panel on Climate Change.

Alonso-Sáez, L., Balagué, V., Sà, E. L., Sánchez, O., González, J. M., Pinhassi, J., ... & Gasol, J. M. (2007). Seasonality in bacterial diversity in north-west Mediterranean coastal waters: assessment through clone libraries, fingerprinting and FISH. FEMS microbiology ecology, 60(1), 98-112.

Alverson, A. J. (2008). Molecular systematics and the diatom species. Protist, 159(3), 339.

Amato, A., Kooistra, W. H. C. F., & Montresor, M. (2018). Cryptic diversity: a long-lasting Issue for diatomologists. Protist, 170(1), 1-7.

Amato, A., Orsini, L., D'Alelio, D., & Montresor, M. (2005). Life cycle, size reduction patterns, and ultrastructure of the pennate planktonic diatom Pseudo-nitzschia delicatissima (Bacillariophyceae) 1. Journal of Phycology, 41(3), 542-556.

Amin, S. A., Hmelo, L. R., Van Tol, H. M., Durham, B. P., Carlson, L. T., Heal, K. R., ... & Armbrust, E. V. (2015). Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria. Nature, 522(7554), 98-101.

Amin, S. A., Parker, M. S., & Armbrust, E. V. (2012). Interactions between diatoms and bacteria. Microbiology and Molecular Biology Reviews, 76(3), 667-684.

Anderson, M. J., Crist, T. O., Chase, J. M., Vellend, M., Inouye, B. D., Freestone, A. L., ... & Swenson, N. G. (2011). Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. Ecology letters, 14(1), 19-28.

Annunziata, R., Mele, B. H., Marotta, P., Volpe, M., Entrambasaguas, L., Mager, S., ... & Ferrante, M. I. (2022). Trade-off between sex and growth in diatoms: Molecular mechanisms and demographic implications. Science advances, 8(3), eabj9466.

Antão, L. I. H. (2019). Effects of ecological scaling on biodiversity patterns. Universidade de Aveiro (Portugal).

Arapov, J., BUŽANČIĆ, M., Penna, A., Casabianca, S., Capellacci, S., Andreoni, F., ... & GLADANA, Ž. N. (2020). High proliferation of Pseudo-nitzschia cf. arenysensis in the Adriatic Sea: ecological and morphological characterisation. Mediterranean Marine Science, 21(3), 759-774.

Ardyna, M., & Arrigo, K. R. (2020). Phytoplankton dynamics in a changing Arctic Ocean. Nature Climate Change, 10(10), 892-903.

Ardyna, M., Babin, M., Gosselin, M., Devred, E., Rainville, L., & Tremblay, J. É. (2014). Recent Arctic Ocean sea ice loss triggers novel fall phytoplankton blooms. Geophysical Research Letters, 41(17), 6207-6212.

Armbrust, E. V. (2009). The life of diatoms in the world's oceans. Nature, 459(7244), 185-192.

Arrhenius, O. (1920). Distribution of species over the area. Medeland. Vedenskaps Akad. Nobel-Inst., 4, 1-6.

Arrhenius, O. (1921). Species and area. Journal of Ecology, 9(1), 95-99.

Arrigo, K. R., & van Dijken, G. L. (2011). Secular trends in Arctic Ocean net primary production. Journal of Geophysical Research: Oceans, 116(C9).

Arrigo, K. R., & van Dijken, G. L. (2015). Continued increases in Arctic Ocean primary production. Progress in Oceanography, 136, 60-70.

Årthun, M., Eldevik, T., Smedsrud, L. H., Skagseth, Ø., & Ingvaldsen, R. B. (2012). Quantifying the influence of Atlantic heat on Barents Sea ice variability and retreat. Journal of Climate, 25(13), 4736-4743.

Assmy, P. et al. Thick-shelled, grazer-protected diatoms decouple ocean carbon and silicon cycles in the iron-limited Antarctic circumpolar current. Proc. Natl Acad. Sci. USA 110, 20633–20638 (2013).;

Aumont, O., Éthé, C., Tagliabue, A., Bopp, L., & Gehlen, M. (2015). PISCES-v2: an ocean biogeochemical model for carbon and ecosystem studies. Geoscientific Model Development, 8(8), 2465-2513.

Azevedo, R. B., & Leroi, A. M. (2001). A power law for cells. Proceedings of the National Academy of Sciences, 98(10), 5699-5704.

Azovsky, A. I. (2000). Concept of scale in marine ecology: linking the words or the worlds?. Web ecology, 1(1), 28-34.

Azovsky, A. I. (2002). Size-dependent species-area relationships in benthos: is the world more diverse for microbes?. Ecography, 25(3), 273-282.

Baas-Becking, L. G. M. (1934). Geobiologie; of inleiding tot de milieukunde. WP Van Stockum & Zoon NV.

Baldisserotto, C., Sabia, A., Guerrini, A., Demaria, S., Maglie, M., Ferroni, L., & Pancaldi, S. (2021). Mixotrophic cultivation of Thalassiosira pseudonana with pure and crude glycerol: Impact on lipid profile. Algal Research, 54, 102194.

Balzano, S., Percopo, I., Siano, R., Gourvil, P., Chanoine, M., Marie, D., ... & Sarno, D. (2017). Morphological and genetic diversity of Beaufort Sea diatoms with high contributions from the Chaetoceros neogracilis species complex. Journal of Phycology, 53(1), 161-187.

Barberán, A., Bates, S. T., Casamayor, E. O., & Fierer, N. (2012). Using network analysis to explore co-occurrence patterns in soil microbial communities. The ISME journal, 6(2), 343-351.

Barton, A. D., Dutkiewicz, S., Flierl, G., Bragg, J., & Follows, M. J. (2010). Patterns of diversity in marine phytoplankton. Science, 327(5972), 1509-1511.

Barton, A. D., Irwin, A. J., Finkel, Z. V., & Stock, C. A. (2016). Anthropogenic climate change drives shift and shuffle in North Atlantic phytoplankton communities. Proceedings of the National Academy of Sciences, 113(11), 2964-2969.

Barton, N. H., & Charlesworth, B. (1998). Why sex and recombination?. Science, 281(5385), 1986-1990.

Baselga, A. (2010). Partitioning the turnover and nestedness components of beta diversity. Global ecology and biogeography, 19(1), 134-143.

Basu, S., Patil, S., Mapleson, D., Russo, M. T., Vitale, L., Fevola, C., ... & Ferrante, M. I. (2017). Finding a partner in the ocean: molecular and evolutionary bases of the response to sexual cues in a planktonic diatom. New Phytologist, 215(1), 140-156.

Bates, S. S., Garrison, D. L., & Horner, R. A. (1998). Bloom dynamics and physiology of domoic-acid-producing Pseudo-nitzschia species. NATO ASI series G ecological sciences, 41, 267-292.

Bates, S. S., Hubbard, K. A., Lundholm, N., Montresor, M., & Leaw, C. P. (2018). Pseudo-nitzschia, Nitzschia, and domoic acid: new research since 2011. Harmful Algae, 79, 3-43.

Bavestrello, G., Arillo, A., Calcinai, B., Cattaneo-Vietti, R., Cerrano, C., Gaino, E., ... & Sara, M. (2000). Parasitic diatoms inside Antarctic sponges. The Biological Bulletin, 198(1), 29-33.

Beaugrand, G., Lindley, J. A., Helaouet, P., & Bonnet, D. (2007). Macroecological study of Centropages typicus in the North Atlantic Ocean. Progress in Oceanography, 72(2-3), 259-273.

Behnke, J., & LaRoche, J. (2020). Iron uptake proteins in algae and the role of Iron Starvation-Induced Proteins (ISIPs). European Journal of Phycology, 55(3), 339-360.

Behrenfeld, M. J., Boss, E. S., & Halsey, K. H. (2021a). Phytoplankton community structuring and succession in a competition-neutral resource landscape. ISME Communications, 1(1), 1-8.

Behrenfeld, M. J., Halsey, K. H., Boss, E., Karp-Boss, L., Milligan, A. J., & Peers, G. (2021c). Thoughts on the evolution and ecological niche of diatoms. Ecological Monographs, 91(3), e01457.

Behrenfeld, M. J., Moore, R. H., Hostetler, C. A., Graff, J., Gaube, P., Russell, L. M., ... & Ziemba, L. (2019). The North Atlantic aerosol and marine ecosystem study (NAAMES): science motive and mission overview. Frontiers in Marine Science, 6, 122.

Behrenfeld, M. J., O'Malley, R., Boss, E., Karp-Boss, L., & Mundt, C. (2021b). Phytoplankton biodiversity and the inverted paradox. ISME Communications, 1(1), 1-9.

Behringer, G., Ochsenkühn, M. A., Fei, C., Fanning, J., Koester, J. A., & Amin, S. A. (2018). Bacterial communities of diatoms display strong conservation across strains and time. Frontiers in microbiology, 9, 659.

Bélanger, S., Cizmeli, S. A., Ehn, J., Matsuoka, A., Doxaran, D., Hooker, S., & Babin, M. (2013). Light absorption and partitioning in Arctic Ocean surface waters: impact of multiyear ice melting. Biogeosciences, 10(10), 6433-6452.

Benedetti, F., Vogt, M., Elizondo, U. H., Righetti, D., Zimmermann, N. E., & Gruber, N. (2021). Major restructuring of marine plankton assemblages under global warming. Nature communications, 12(1), 1-15.

Berger, W. H., & Parker, F. L. (1970). Diversity of planktonic foraminifera in deep-sea sediments. Science, 168(3937), 1345-1347.

Beszczynska-Möller, A., Fahrbach, E., Schauer, U., & Hansen, E. (2012). Variability in Atlantic water temperature and transport at the entrance to the Arctic Ocean, 1997–2010. ICES Journal of Marine Science, 69(5), 852-863.

Bilcke, G. (2021). Life history regulation of the benthic pennate diatom Seminavis robusta: a transcriptomic, comparative genomic and physiological study (Doctoral dissertation, Ghent University).

Bilcke, G., Van den Berge, K., De Decker, S., Bonneure, E., Poulsen, N., Bulankova, P., ... & Vyverman, W. (2021). Mating type specific transcriptomic response to sex inducing pheromone in the pennate diatom Seminavis robusta. The ISME journal, 15(2), 562-576.

Blasco-Moreno, A., Pérez-Casany, M., Puig, P., Morante, M., & Castells, E. (2019). What does a zero mean? Understanding false, random and structural zeros in ecology. Methods in Ecology and Evolution, 10(7), 949-959.

Blatch, G. L., & Lässle, M. (1999). The tetratricopeptide repeat: a structural motif mediating protein-protein interactions. Bioessays, 21(11), 932-939.

Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., ... & De Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. Trends in ecology & evolution, 29(6), 358-367.

Bolwell, G. P., Callow, J. A., Callow, M. E., & Evans, L. V. (1979). Fertilization in brown algae. II. Evidence for lectin-sensitive complementary receptors involved in gamete recognition in Fucus serratus. Journal of Cell Science, 36(1), 19-30.

Booth, B. C., & Horner, R. A. (1997). Microalgae on the Arctic Ocean Section, 1994: species abundance and biomass. Deep Sea Research Part II: Topical Studies in Oceanography, 44(8), 1607-1622.

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. Geophysical Research Letters, 32(19).

Bopp, L., Resplandy, L., Orr, J. C., Doney, S. C., Dunne, J. P., Gehlen, M., ... & Vichi, M. (2013). Multiple stressors of ocean ecosystems in the 21st century: projections with CMIP5 models. Biogeosciences, 10(10), 6225-6245.

Borgnino, M., Arrieta, J., Boffetta, G., De Lillo, F., & Tuval, I. (2019). Turbulence induces clustering and segregation of non-motile, buoyancy-regulating phytoplankton. Journal of the Royal Society Interface, 16(159), 20190324.

Borgonuovo, C. (2019). Exploring the sexual phase of the diatom Pseudo-nitzschia multistriata: from genes to metabolites (Doctoral dissertation, The Open University).

Borgonuovo, C., Campese, L., Bilcke, G., Annunziata, R., Van den Berge, K., Montresor, M., Vandepoele, K., Vyverman, W., Iudicone, D., Ferrante, M. I. Meta-omics to explore a panel of molecular markers for the detection of sex events at sea in diatoms (in prep.)

Boscaro, V., Santoferrara, L. F., Zhang, Q., Gentekaki, E., Syberg-Olsen, M. J., Del Campo, J., & Keeling, P. J. (2018). EukRef-Ciliophora: a manually curated, phylogeny-based database of small subunit rRNA gene sequences of ciliates. Environmental microbiology, 20(6), 2218-2230.

Bowler, C., Vardi, A., & Allen, A. E. (2010). Oceanographic and biogeochemical insights from diatom genomes. Annual review of marine science, 2, 333-365.

Boyd, P. W., & Ellwood, M. J. (2010). The biogeochemical cycle of iron in the ocean. Nature Geoscience, 3(10), 675-682.

Boyd, P. W., Collins, S., Dupont, S., Fabricius, K., Gattuso, J. P., Havenhand, J., ... & Pörtner, H. O. (2018). Experimental strategies to assess the biological ramifications of multiple drivers of global ocean change—a review. Global change biology, 24(6), 2239-2261.

Boyer, T. P., Baranova, O. K., Coleman, C., Garcia, H. E., Grodsky, A., Locarnini, R. A., ... & Zweng, M. M. (2018). NOAA Atlas NESDIS 87. World Ocean Database.

Boyer, T. P.; Garcia, H. E.; Locarnini, R. A.; Zweng, M. M.; Mishonov, A. V.; Reagan, J. R.; Weathers, K. A.; Baranova, O. K.; Seidov, D.; Smolyar, I. V. (2018). World Ocean Atlas 2018. NOAA National Centers for Environmental Information. Dataset. https://accession.nodc.noaa.gov/NCEI-WOA18.

Bracco, A., Provenzale, A., & Scheuring, I. (2000). Mesoscale vortices and the paradox of the plankton. Proceedings of the Royal Society of London. Series B: Biological Sciences, 267(1454), 1795-1800.

Brock, T., 1987. The study of microorganisms in-situ: progress and problems. In: Fletcher, M., Gray, T., Jones, J. (Eds.), Cambridge University Press, pp. 1–17.

Brown, J. H. (1995). Macroecology. Chicago: Univ.

Brown, S. P., Veach, A. M., Rigdon-Huss, A. R., Grond, K., Lickteig, S. K., Lothamer, K., ... & Jumpponen, A. (2015). Scraping the bottom of the barrel: are rare high throughput sequences artifacts?. Fungal Ecology, 13, 221-225.

Browning, T. J., Al-Hashem, A. A., Hopwood, M. J., Engel, A., Wakefield, E. D., & Achterberg, E. P. (2020). Nutrient regulation of late spring phytoplankton blooms in the midlatitude North Atlantic. Limnology and Oceanography, 65(6), 1136-1148.

Brunson, J. K., McKinnie, S. M., Chekan, J. R., McCrow, J. P., Miles, Z. D., Bertrand, E. M., ... & Moore, B. S. (2018). Biosynthesis of the neurotoxin domoic acid in a bloom-forming diatom. Science, 361(6409), 1356-1358.

Bucchini, F., Del Cortona, A., Kreft, Ł., Botzki, A., Van Bel, M., & Vandepoele, K. (2021). TRAPID 2.0: a web application for taxonomic and functional analysis of de novo transcriptomes. Nucleic acids research, 49(17), e101-e101.

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

Bulankova, P., Sekulic, M., Jallet, D., Nef, C., Delmont, T. O., van Oosterhout, C., ... & De Veylder, L. (2020). Mitotic interhomolog recombination drives genomic diversity in diatoms. bioRxiv.

Burki, F., Sandin, M. M., & Jamy, M. (2021). Diversity and ecology of protists revealed by metabarcoding. Current Biology, 31(19), R1267-R1280.

Busseni, G., Caputi, L., Piredda, R., Fremont, P., Hay Mele, B., Campese, L., ... & Iudicone, D. (2020). Large scale patterns of marine diatom richness: Drivers and trends in a changing ocean. Global Ecology and Biogeography, 29(11), 1915-1928.

Busseni, G., Rocha Jimenez Vieira, F., Amato, A., Pelletier, E., Pierella Karlusich, J. J., Ferrante, M. I., ... & Iudicone, D. (2019). Meta-omics reveals genetic flexibility of diatom nitrogen transporters in response to environmental changes. Molecular biology and evolution, 36(11), 2522-2535.

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME journal, 11(12), 2639-2643.

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics, 25(15), 1972-1973.

Caputi, L., Carradec, Q., Eveillard, D., Kirilovsky, A., Pelletier, E., Pierella Karlusich, J. J., ... & Tara Oceans Coordinators. (2019). Community-level responses to iron availability in open ocean plankton ecosystems. Global Biogeochemical Cycles, 33(3), 391-419.

Caron, D. A., & Countway, P. D. (2009). Hypotheses on the role of the protistan rare biosphere in a changing world. Aquatic Microbial Ecology, 57(3), 227-238.

Carr, M. H., Neigel, J. E., Estes, J. A., Andelman, S., Warner, R. R., & Largier, J. L. (2003). Comparing marine and terrestrial ecosystems: implications for the design of coastal marine reserves. Ecological Applications, 13(sp1), 90-107.

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., ... & Wincker, P. (2018). A global ocean atlas of eukaryotic genes. Nature communications, 9(1), 1-13.

Casteleyn, G., Leliaert, F., Backeljau, T., Debeer, A. E., Kotaki, Y., Rhodes, L., ... & Vyverman, W. (2010). Limits to gene flow in a cosmopolitan marine planktonic diatom. Proceedings of the National Academy of Sciences, 201001380.

Cavole, L. M., Demko, A. M., Diner, R. E., Giddings, A., Koester, I., Pagniello, C. M., ... & Franks, P. J. (2016). Biological impacts of the 2013–2015 warm-water anomaly in the Northeast Pacific: winners, losers, and the future. Oceanography, 29(2), 273-285.

Cerino, F., Orsini, L., Sarno, D., Dell'Aversano, C., Tartaglione, L., & Zingone, A. (2005). The alternation of different morphotypes in the seasonal cycle of the toxic diatom Pseudo-nitzschia galaxiae. Harmful Algae, 4(1), 33-48.

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. Journal of Phycology, 49(6), 1128-1141.

Chao, A., & Jost, L. (2012). Diversity measures. In Encyclopedia of theoretical ecology (pp. 203-207). University of California Press.

Chepurnov, V. A., & Mann, D. G. (2004). Auxosporulation of Licmophora communis (Bacillariophyta) and a review of mating systems and sexual reproduction in araphid pennate diatoms. Phycological Research, 52(1), 1-12.

Chepurnov, V. A., Mann, D. G., Sabbe, K., Vannerum, K., Casteleyn, G., Verleyen, E., ... & Vyverman, W. (2005). Sexual reproduction, mating system, chloroplast dynamics and abrupt cell size reduction in Pseudo-nitzschia pungens from the North Sea (Bacillariophyta). European Journal of Phycology, 40*(4), 379-395.*

Chisholm, R. A., & Pacala, S. W. (2010). Niche and neutral models predict asymptotically equivalent species abundance distributions in high-diversity ecological communities. Proceedings of the National Academy of Sciences, 107(36), 15821-15825.

Chivers, W. J., Walne, A. W., & Hays, G. C. (2017). Mismatch between marine plankton range movements and the velocity of climate change. Nature Communications, 8(1), 1-8.

Chust, G., Irigoien, X., Chave, J., & Harris, R. P. (2013). Latitudinal phytoplankton distribution and the neutral theory of biodiversity. Global Ecology and Biogeography, 22(5), 531-543.

Ciais, P., Sabine, C., Bala, G., Bopp, L., Brovkin, V., Canadell, J., ... & Thornton, P. (2014). Carbon and other biogeochemical cycles. In Climate change 2013: the physical science basis. Contribution of Working Group

I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (pp. 465-570). Cambridge University Press.

Clark, G. F. (2011). Molecular models for mouse sperm-oocyte binding. Glycobiology, 21(1), 3-5.

Clark, S., Hubbard, K. A., McGillicuddy Jr, D. J., Ralston, D. K., & Shankar, S. (2021). Investigating Pseudo-nitzschia australis introduction to the Gulf of Maine with observations and models. Continental Shelf Research, 228, 104493.

Codispoti, L. A., Kelly, V., Thessen, A., Matrai, P., Suttles, S., Hill, V., ... & Light, B. (2013). Synthesis of primary production in the Arctic Ocean: III. Nitrate and phosphate based estimates of net community production. Progress in Oceanography, 110, 126-150.

Cohen, J. E., & Xu, M. (2015). Random sampling of skewed distributions implies Taylor's power law of fluctuation scaling. Proceedings of the National Academy of Sciences, 112(25), 7749-7754.

Cohen, N. R., Ellis, K. A., Lampe, R. H., McNair, H., Twining, B. S., Maldonado, M. T., ... & Marchetti, A. (2017). Diatom transcriptional and physiological responses to changes in iron bioavailability across ocean provinces. Frontiers in Marine Science, 4, 360.

Cota, G. F. (1985). Photoadaptation of high Arctic ice algae. Nature, 315(6016), 219-222.

Council, A. (2015). Conservation of Arctic Flora and Fauna (CAFF),". CAFF," available at arctic-council. org/working_group/caff.

Coyne, J. A., & Orr, H. A. (2004). Speciation (Vol. 37). Sunderland, MA: Sinauer Associates.

Crossette, E., Gumm, J., Langenfeld, K., Raskin, L., Duhaime, M., & Wigginton, K. (2021). Metagenomic quantification of genes with internal standards. MBio, 12(1), e03173-20.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. InterJournal, complex systems, 1695(5), 1-9.

D'Alelio, D., d'Alcala, M. R., Dubroca, L., Zingone, A., & Montresor, M. (2010). The time for sex: a biennial life cycle in a marine planktonic diatom. Limnology and Oceanography, 55(1), 106-114.

Davidovich, N. A., & Bates, S. S. (1998). Sexual reproduction in the pennate diatoms Pseudo-nitzschia multiseries and P. pseudodelicatissima (Bacillariophyceae). Journal of Phycology, 34(1), 126-137.

De Luca, D., Kooistra, W. H., Sarno, D., Gaonkar, C. C., & Piredda, R. (2019). Global distribution and diversity of Chaetoceros (Bacillariophyta, Mediophyceae): integration of classical and novel strategies. PeerJ, 7, e7410.

De Luca, D., Piredda, R., Sarno, D., & Kooistra, W. H. (2021). Resolving cryptic species complexes in marine protists: phylogenetic haplotype networks meet global DNA metabarcoding datasets. The ISME Journal, 15(7), 1931-1942.

De Steur, L., Hansen, E., Mauritzen, C., Beszczynska-Möller, A., & Fahrbach, E. (2014). Impact of recirculation on the East Greenland Current in Fram Strait: Results from moored current meter measurements between 1997 and 2009. Deep Sea Research Part I: Oceanographic Research Papers, 92, 26-40.

De Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., ... & Carmichael, M. (2015). Eukaryotic plankton diversity in the sunlit ocean. Science, 348(6237).

Del Campo, J., Kolisko, M., Boscaro, V., Santoferrara, L. F., Nenarokov, S., Massana, R., ... & Wegener Parfrey, L. (2018). EukRef: phylogenetic curation of ribosomal RNA to enhance understanding of eukaryotic diversity and distribution. PLoS biology, 16(9), e2005849.

Delmont, T. O., Gaia, M., Hinsinger, D. D., Frémont, P., Vanni, C., Fernandez-Guerra, A., ... & Pelletier, E. (2022). Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. Cell Genomics, 2(5), 100123.

Di Lorenzo, E., & Mantua, N. (2016). Multi-year persistence of the 2014/15 North Pacific marine heatwave. Nature Climate Change, 6(11), 1042-1047.

Díaz, S., Fargione, J., Chapin III, F. S., & Tilman, D. (2006). Biodiversity loss threatens human well-being. PLoS biology, 4(8), e277.

Díaz, S., Purvis, A., Cornelissen, J. H., Mace, G. M., Donoghue, M. J., Ewers, R. M., ... & Pearse, W. D. (2013). Functional traits, the phylogeny of function, and ecosystem service vulnerability. Ecology and evolution, 3(9), 2958-2975.

Doney, S. C., Ruckelshaus, M., Emmett Duffy, J., Barry, J. P., Chan, F., English, C. A., ... & Talley, L. D. (2012). Climate change impacts on marine ecosystems. Annual review of marine science, 4, 11-37.

Dong, H. C., Lundholm, N., Teng, S. T., Li, A., Wang, C., Hu, Y., & Li, Y. (2020). Occurrence of Pseudo-nitzschia species and associated domoic acid production along the Guangdong coast, South China Sea. Harmful algae, 98, 101899.

Dore, J. E., Lukas, R., Sadler, D. W., Church, M. J., & Karl, D. M. (2009). Physical and biogeochemical modulation of ocean acidification in the central North Pacific. Proceedings of the National Academy of Sciences, 106(30), 12235-12240.

Drakare, S., Lennon, J. J., & Hillebrand, H. (2006). The imprint of the geographical, evolutionary and ecological context on species–area relationships. Ecology letters, 9(2), 215-227.

Duarte, C. M. (2015). Seafaring in the 21st century: the Malaspina 2010 circumnavigation expedition. Limnology and Oceanography Bulletin, 24(1), 11-14.

Dusa, A. (2020). venn: Draw Venn Diagrams. R Packag. version 1.9.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC bioinformatics, 5(1), 1-19.

Egge, E., Elferink, S., Vaulot, D., John, U., Bratbak, G., Larsen, A., & Edvardsen, B. (2021). An 18S V4 rRNA metabarcoding dataset of protist diversity in the Atlantic inflow to the Arctic Ocean, through the year and down to 1000 m depth. Earth System Science Data, 13(10), 4913-4928.

Eiler, A., Heinrich, F., & Bertilsson, S. (2012). Coherent dynamics and association networks among lake bacterioplankton taxa. The ISME journal, 6(2), 330-342.

Elbrecht, V., Vamos, E. E., Steinke, D., & Leese, F. (2018). Estimating intraspecific genetic diversity from community DNA metabarcoding data. PeerJ, 6, e4644.

Endo, H., Blanc-Mathieu, R., Li, Y., Salazar, G., Henry, N., Labadie, K., ... & Ogata, H. (2020). Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. Nature Ecology & Evolution, 4(12), 1639-1649.

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic acids research, 30(7), 1575-1584.

Erdner, D. L., Richlen, M., McCauley, L. A., & Anderson, D. M. (2011). Diversity and dynamics of a widespread bloom of the toxic dinoflagellate Alexandrium fundyense. PloS one, 6(7), e22965.

Espinoza, J. L., Shah, N., Singh, S., Nelson, K. E., & Dupont, C. L. (2020). Applications of weighted association networks applied to compositional data in biology. Environmental Microbiology, 22(8), 3020-3038.

Estrada, M., Delgado, M., Blasco, D., Latasa, M., Cabello, A. M., Benítez-Barrios, V., ... & Vidal, M. (2016). Phytoplankton across tropical and subtropical regions of the Atlantic, Indian and Pacific oceans. PLoS One, 11(3), e0151699.

Etienne, R. S., & Olff, H. (2005). Confronting different models of community structure to species-abundance data: a Bayesian model comparison. Ecology letters, 8(5), 493-504.

Euskirchen, E. S., Goodstein, E. S., & Huntington, H. P. (2013). An estimated cost of lost climate regulation services caused by thawing of the Arctic cryosphere. Ecological applications, 23(8), 1869-1880.

Falciatore, A., Jaubert, M., Bouly, J. P., Bailleul, B., & Mock, T. (2020). Diatom molecular research comes of age: model species for studying phytoplankton biology and diversity. The Plant Cell, 32(3), 547-572.

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

Falkowski, P. G., & Knoll, A. H. (2007). An introduction to primary producers in the sea: who they are, what they do, and when they evolved. In Evolution of primary producers in the sea (pp. 1-6). Academic Press.

Falkowski, P. G., Katz, M. E., Knoll, A. H., Quigg, A., Raven, J. A., Schofield, O., & Taylor, F. J. R. (2004). The evolution of modern eukaryotic phytoplankton. science, 305(5682), 354-360.

Faure, E., Not, F., Benoiston, A. S., Labadie, K., Bittner, L., & Ayata, S. D. (2019). Mixotrophic protists display contrasted biogeographies in the global ocean. The ISME journal, 13(4), 1072-1083.

Faust, K., Lima-Mendez, G., Lerat, J. S., Sathirapongsasuti, J. F., Knight, R., Huttenhower, C., ... & Raes, J. (2015). Cross-biome comparison of microbial association networks. Frontiers in microbiology, 6, 1200.

Fauth, J. E., Bernardo, J., Camara, M., Resetarits Jr, W. J., Van Buskirk, J., & McCollum, S. A. (1996). Simplifying the jargon of community ecology: a conceptual approach. The American Naturalist, 147(2), 282-286.

Fenchel, T., Esteban, G. F., & Finlay, B. J. (1997). Local versus global diversity of microorganisms: cryptic diversity of ciliated protozoa. Oikos, 220-225.

Ferrante, M. I., Entrambasaguas, L., Johansson, M., Töpel, M., Kremp, A., Montresor, M., & Godhe, A. (2019). Exploring molecular signs of sex in the marine diatom Skeletonema marinoi. Genes, 10(7), 494.

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

Fierer, N., & Jackson, R. B. (2006). The diversity and biogeography of soil bacterial communities. Proceedings of the National Academy of Sciences, 103(3), 626-631.

Finlay, B. J. (2002). Global dispersal of free-living microbial eukaryote species. Science, 296(5570), 1061-1063.

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic acids research, 39(suppl_2), W29-W37.

Fisher, R. A., Corbet, A. S., & Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. The Journal of Animal Ecology, 42-58.

Fitzpatrick, E., Caron, D. A., & Schnetzer, A. (2010). Development and environmental application of a genus-specific quantitative PCR approach for Pseudo-nitzschia species. Marine biology, 157(5), 1161-1169.

Flegontova, O., Flegontov, P., Malviya, S., Audic, S., Wincker, P., De Vargas, C., ... & Horák, A. (2016). Extreme diversity of diplonemid eukaryotes in the ocean. Current Biology, 26(22), 3060-3065.

Flynn, K. J., Mitra, A., Anestis, K., Anschütz, A. A., Calbet, A., Ferreira, G. D., ... & Traboni, C. (2019). Mixotrophic protists and a new paradigm for marine ecology: where does plankton research go now?. Journal of Plankton Research, 41(4), 375-391.

Foissner, W. (2006). Biogeography and dispersal of micro-organisms: a review emphasizing protists. Acta protozoologica, 45(2), 111-136.

Forster, D., Lentendu, G., Wilson, M., Mahé, F., Leese, F., Andersen, T., ... & Dunthorn, M. (2020). Evaluating geographic variation within molecular operational taxonomic units (OTUs) using network analyses in Scandinavian lakes. BioRxiv.

Foster, Z. S., Sharpton, T. J., & Grünwald, N. J. (2017). Metacoder: An R package for visualization and manipulation of community taxonomic diversity data. PLoS computational biology, 13(2), e1005404.

Gabaldón, T., & Koonin, E. V. (2013). Functional and evolutionary implications of gene orthology. Nature Reviews Genetics, 14(5), 360-366.

Galhardo, R. S., Hastings, P. J., & Rosenberg, S. M. (2007). Mutation as a stress response and the regulation of evolvability. Critical reviews in biochemistry and molecular biology, 42(5), 399-435.

Gao, X., Bowler, C., & Kazamia, E. (2021). Iron metabolism strategies in diatoms. Journal of Experimental Botany, 72(6), 2165-2180.

Gaonkar, C. C., Kooistra, W. H., Lange, C. B., Montresor, M., & Sarno, D. (2017). Two new species in the Chaetoceros socialis complex (Bacillariophyta): C. sporotruncatus and C. dichatoensis, and characterization of its relatives, C. radicans and C. cinctus. Journal of phycology, 53(4), 889-907.

Gaonkar, C. C., Piredda, R., Minucci, C., Mann, D. G., Montresor, M., Sarno, D., & Kooistra, W. H. (2018). Annotated 18S and 28S rDNA reference sequences of taxa in the planktonic diatom family Chaetocerotaceae. PloS one, 13(12), e0208929.

Gaonkar, C. C., Piredda, R., Sarno, D., Zingone, A., Montresor, M., & Kooistra, W. H. (2020). Species detection and delineation in the marine planktonic diatoms Chaetoceros and Bacteriastrum through metabarcoding: making biological sense of haplotype diversity. Environmental Microbiology, 22(5), 1917-1929.

Garali, S. M. B., Sahraoui, I., Othman, H. B., Kouki, A., de La Iglesia, P., Diogène, J., ... & Hlaili, A. S. (2021). Capacity of the potentially toxic diatoms Pseudo-nitzschia mannii and Pseudo-nitzschia hasleana to tolerate polycyclic aromatic hydrocarbons. Ecotoxicology and Environmental Safety, 214, 112082.

Gaston, K. J., Blackburn, T. M., Greenwood, J. J., Gregory, R. D., Quinn, R. M., & Lawton, J. H. (2000). Abundance–occupancy relationships. Journal of Applied Ecology, 37, 39-59.

Gause, G. F. (1934). Experimental analysis of Vito Volterra's mathematical theory of the struggle for existence. Science, 79(2036), 16-17.

Gehlenborg, N. (2019). UpSetR: a more scalable alternative to Venn and Euler diagrams for visualizing intersecting sets. R package version, 1(0).

Ghiglione, J. F., Mevel, G., Pujo-Pay, M., Mousseau, L., Lebaron, P., & Goutx, M. (2007). Diel and seasonal variations in abundance, activity, and community structure of particle-attached and free-living bacteria in NW Mediterranean Sea. Microbial ecology, 54(2), 217-231.

Gilbert, B., & Lechowicz, M. J. (2004). Neutrality, niches, and dispersal in a temperate forest understory. Proceedings of the National Academy of Sciences, 101(20), 7651-7656.

Gillard, J., Frenkel, J., Devos, V., Sabbe, K., Paul, C., Rempt, M., ... & Vyverman, W. (2013). Metabolomics enables the structure elucidation of a diatom sex pheromone. Angewandte Chemie International Edition, 52(3), 854-857.

Giulietti, S., Romagnoli, T., Siracusa, M., Bacchiocchi, S., Totti, C., & Accoroni, S. (2021). Integrative taxonomy of the Pseudo-nitzschia (Bacillariophyceae) populations in the NW Adriatic Sea, with a focus on a novel cryptic species in the P. delicatissima species complex. Phycologia, 60(3), 247-264.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. Frontiers in microbiology, 8, 2224.

Godhe, A., & Härnström, K. (2010). Linking the planktonic and benthic habitat: genetic structure of the marine diatom Skeletonema marinoi. Molecular Ecology, 19(20), 4478-4490.

Godhe, A., & Rynearson, T. (2017). The role of intraspecific variation in the ecological and evolutionary success of diatoms in changing environments. Philosophical Transactions of the Royal Society B: Biological Sciences, 372(1728), 20160399.

Godhe, A., Sjöqvist, C., Sildever, S., Sefbom, J., Harðardóttir, S., Bertos-Fortis, M., ... & Khandan, S. (2016). Physical barriers and environmental gradients cause spatial and temporal genetic differentiation of an extensive algal bloom. Journal of Biogeography, 43(6), 1130-1142.

Godon, J. J., Arulazhagan, P., Steyer, J. P., & Hamelin, J. (2016). Vertebrate bacterial gut diversity: size also matters. BMC ecology, 16(1), 12.

Goodenough, U., Lin, H., & Lee, J. H. (2007, June). Sex determination in Chlamydomonas. In Seminars in cell & developmental biology (Vol. 18, No. 3, pp. 350-361). Academic Press.

Gray, J. S., Bjørgesæter, A., Ugland, K. I., & Frank, K. (2006). Are there differences in structure between marine and terrestrial assemblages?. Journal of experimental marine biology and ecology, 330(1), 19-26.

Green, J. L., Holmes, A. J., Westoby, M., Oliver, I., Briscoe, D., Dangerfield, M., ... & Beattie, A. J. (2004). Spatial scaling of microbial eukaryote diversity. Nature, 432(7018), 747-750.

Gregory, A. C., Zayed, A. A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., ... & Roux, S. (2019). Marine DNA viral macro-and microdiversity from pole to pole. Cell, 177(5), 1109-1123.

Grey, E. K., Bernatchez, L., Cassey, P., Deiner, K., Deveney, M., Howland, K. L., ... & Lodge, D. M. (2018). Effects of sampling effort on biodiversity patterns estimated from environmental DNA metabarcoding surveys. Scientific Reports, 8(1), 1-10.

Grilli, J. (2020). Macroecological laws describe variation and diversity in microbial communities. Nature communications, 11(1), 1-11.

Grishkovskaya, I., Avvakumov, G. V., Sklenar, G., Dales, D., Hammond, G. L., & Muller, Y. A. (2000). Crystal structure of human sex hormone-binding globulin: steroid transport by a laminin G-like domain. The EMBO journal, 19(4), 504-512.

Guangchuang, Y. (2020). scatterpie: Scatter Pie Plot. R package version 0.1, 5.

Guannel, M. L., Horner-Devine, M. C., & Rocap, G. (2011). Bacterial community composition differs with species and toxigenicity of the diatom Pseudo-nitzschia. Aquatic Microbial Ecology, 64(2), 117-133.

Guihéneuf, F., Mimouni, V., Ulmann, L., & Tremblin, G. (2008). Environmental factors affecting growth and omega 3 fatty acid composition in Skeletonema costatum. The influences of irradiance and carbon source: Communication presented at the 25ème Congrès Annuel de l'Association des Diatomistes de Langue Francaise (ADLaF), Caen, 25-28 September 2006. Diatom Research, 23(1), 93-103.

Guillard, R., & Kilham, P. (1977). The ecology of marine planktonic diatoms. The biology of diatoms, 13, 372-469.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., ... & Christen, R. (2012). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. Nucleic acids research, 41(D1), D597-D604.

Guiry, M. D. (2012). How many species of algae are there?. Journal of phycology, 48(5), 1057-1063.

Guiry, M.D. & Guiry, G.M. (2021). AlgaeBase. World-wide electronic publication, National University of Ireland, Galway. https://www.algaebase.org

Haine, T. W., Curry, B., Gerdes, R., Hansen, E., Karcher, M., Lee, C., ... & Woodgate, R. (2015). Arctic freshwater export: Status, mechanisms, and prospects. Global and Planetary Change, 125, 13-35.

Hallegraeff, G. M. (2010). Ocean climate change, phytoplankton community responses, and harmful algal blooms: a formidable predictive challenge 1. Journal of phycology, 46(2), 220-235.

Hamilton, N. (2016). ggtern: An Extension to 'ggplot2', for the Creation of Ternary Diagrams. R package version, 2(1).

Hamm, C. E., Merkel, R., Springer, O., Jurkojc, P., Maier, C., Prechtel, K., & Smetacek, V. (2003). Architecture and material properties of diatom shells provide effective mechanical protection. Nature, 421(6925), 841-843.

Hansen, B., Østerhus, S., Turrell,W. R., Jónsson, S., Valdimarsson, H., Hátún, H., & Olsen, S.M. (2008). The inflow of Atlantic water, heat, and salt to the Nordic seas across the Greenland–Scotland ridge, Arctic–Subarctic ocean fluxes (pp. 15–43). Dordrecht: Springer.

Hardin, G. (1960). The competitive exclusion principle. science, 131(3409), 1292-1297.

Hasle, G. R. (2002). Are most of the domoic acid-producing species of the diatom genus Pseudo-nitzschia cosmopolites?. Harmful algae, 1(2), 137-146.

Hays, G. C., Richardson, A. J., & Robinson, C. (2005). Climate change and marine plankton. Trends in ecology & evolution, 20(6), 337-344.

Hebenstreit, D., Fang, M., Gu, M., Charoensawan, V., van Oudenaarden, A., & Teichmann, S. A. (2011). RNA sequencing reveals two major classes of gene expression levels in metazoan cells. Molecular systems biology, 7(1), 497.

Hellweger, F. L., van Sebille, E., & Fredrick, N. D. (2014). Biogeographic patterns in ocean microbes emerge in a neutral agent-based model. Science, 345(6202), 1346-1349.

Hennon, G. M., & Dyhrman, S. T. (2020). Progress and promise of omics for predicting the impacts of climate change on harmful algal blooms. Harmful Algae, 91, 101587.

Hillebrand, H. (2004). On the generality of the latitudinal diversity gradient. The American Naturalist, 163(2), 192-211.

Hollibaugh, J. T., Wong, P. S., & Murrell, M. C. (2000). Similarity of particle-associated and free-living bacterial communities in northern San Francisco Bay, California. Aquatic Microbial Ecology, 21(2), 103-114.

Holmes, R. M., McClelland, J. W., Peterson, B. J., Tank, S. E., Bulygina, E., Eglinton, T. I., ... & Zimov, S. A. (2012). Seasonal and annual fluxes of nutrients and organic matter from large rivers to the Arctic Ocean and surrounding seas. Estuaries and Coasts, 35(2), 369-382.

Horner-Devine, M. C., Lage, M., Hughes, J. B., & Bohannan, B. J. (2004). A taxa–area relationship for bacteria. Nature, 432(7018), 750-753.

Hubbell, S. P. (1997). A unified theory of biogeography and relative species abundance and its application to tropical rain forests and coral reefs. Coral reefs, 16(1), S9-S21.

Hubbell, S. P. (2001). A Unified Theory of Biodiversity and Biogeography–Princeton University Press. Princeton, NJ.

Huisman, J., & Weissing, F. J. (2000). reply: Coexistence and resource competition. Nature, 407(6805), 694-694.

Hutchins, D. A., & Boyd, P. W. (2016). Marine phytoplankton and the changing ocean iron cycle. Nature Climate Change, 6(12), 1072-1079.

Hutchins, D. A., & Bruland, K. W. (1998). Iron-limited diatom growth and Si: N uptake ratios in a coastal upwelling regime. Nature, 393(6685), 561-564.

Hutchins, D. A., & Fu, F. (2017). Microorganisms and ocean global change. Nature microbiology, 2(6), 1-11.

Hutchinson, G. E. (1961). The paradox of the plankton. The American Naturalist, 95(882), 137-145.

Ibarbalz, F. M., Henry, N., Brandão, M. C., Martini, S., Busseni, G., Byrne, H., ... & Zinger, L. (2019). Global trends in marine plankton diversity across kingdoms of life. Cell, 179(5), 1084-1097.

Iglesias-Rodriguez, M. D., Schofield, O. M., Batley, J., Medlin, L. K., & Hayes, P. K. (2006). Intraspecific genetic diversity in the marine coccolithophore Emiliania huxleyi (Prymnesiophyceae): the use of microsatellite analysis in marine phytoplankton population studies. J. Phycol, 42(3), 526-536.

Ingvaldsen, R., Loeng, H., & Asplin, L. (2002). Variability in the Atlantic inflow to the Barents Sea based on a one-year time series from moored current meters. Continental Shelf Research, 22(3), 505-519.

Jakobsson, M. (2002). Hypsometry and volume of the Arctic Ocean and its constituent seas. Geochemistry, Geophysics, Geosystems, 3(5), 1-18.

Jakobsson, M., & Macnab, R. (2006). A comparison between GEBCO sheet 5.17 and the International Bathymetric Chart of the Arctic Ocean (IBCAO) version 1.0. Marine Geophysical Researches, 27(1), 35-48.

Jia, X., Dini-Andreote, F., & Salles, J. F. (2018). Community assembly processes of the microbial rare biosphere. Trends in microbiology, 26(9), 738-747.

Jin, P., & Agustí, S. (2018). Fast adaptation of tropical diatoms to increased warming with trade-offs. Scientific Reports, 8(1), 1-10.

Jin, X., Gruber, N., Dunne, J. P., Sarmiento, J. L., & Armstrong, R. A. (2006). Diagnosing the contribution of phytoplankton functional groups to the production and export of particulate organic carbon, CaCO3, and opal from global nutrient and alkalinity distributions. Global Biogeochemical Cycles, 20(2).

Jungblut, S., Liebich, V., & Bode, M. (2018). YOUMARES 8–Oceans Across Boundaries: Learning from each other: Proceedings of the 2017 conference for YOUng MARine RESearchers in Kiel, Germany (p. 251). Springer Nature.

Kahla, O., Garali, S. M. B., Karray, F., Abdallah, M. B., Kallel, N., Mhiri, N., ... & Hlaili, A. S. (2021). Efficiency of benthic diatom-associated bacteria in the removal of benzo (a) pyrene and fluoranthene. Science of the Total Environment, 751, 141399.

Kahru, M., Brotas, V., Manzano-Sarabia, M., & Mitchell, B. G. (2011). Are phytoplankton blooms occurring earlier in the Arctic?. Global Change Biology, 17(4), 1733-1739.

Karsenti, E., Acinas, S. G., Bork, P., Bowler, C., De Vargas, C., Raes, J., ... & Tara Oceans Consortium. (2011). A holistic approach to marine eco-systems biology. PLoS biology, 9(10), e1001177.

Kazamia, E., Sutak, R., Paz-Yepes, J., Dorrell, R. G., Vieira, F. R. J., Mach, J., ... & Lesuisse, E. (2018). Endocytosis-mediated siderophore uptake as a strategy for Fe acquisition in diatoms. Science Advances, 4(5), eaar4536.

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., ... & Worden, A. Z. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS biology, 12(6), e1001889.

Kemp, A. E., & Villareal, T. A. (2018). The case of the diatoms and the muddled mandalas: Time to recognize diatom adaptations to stratified waters. Progress in Oceanography, 167, 138-149.

Khavari, F., Saidijam, M., Taheri, M., & Nouri, F. (2021). Microalgae: therapeutic potentials and applications. Molecular Biology Reports, 1-9.

Kieft, T. L. (2017). New allometric scaling laws revealed for microorganisms. Trends in Ecology & Evolution, 32(6), 400-402.

Kim, J. H., Ajani, P., Murray, S. A., Kim, J. H., Lim, H. C., Teng, S. T., ... & Park, B. S. (2020). Sexual reproduction and genetic polymorphism within the cosmopolitan marine diatom Pseudo-nitzschia pungens. Scientific reports, 10(1), 1-13.

Kim, J. H., Wang, P., Park, B. S., Kim, J. H., Patidar, S. K., & Han, M. S. (2018). Revealing the distinct habitat ranges and hybrid zone of genetic sub-populations within Pseudo-nitzschia pungens (Bacillariophyceae) in the West Pacific area. Harmful algae, 73, 72-83.

Koleff, P., Gaston, K. J., & Lennon, J. J. (2003). Measuring beta diversity for presence–absence data. Journal of Animal Ecology, 72(3), 367-382.

Kondratyeva, A., Grandcolas, P., & Pavoine, S. (2019). Reconciling the concepts and measures of diversity, rarity and originality in ecology and evolution. Biological Reviews, 94(4), 1317-1337.

Kooistra, W. H., & Medlin, L. K. (1996). Evolution of the diatoms (Bacillariophyta): IV. A reconstruction of their age from small subunit rRNA coding regions and the fossil record. Molecular phylogenetics and evolution, 6(3), 391-407.

Kooistra, W. H., Gersonde, R., Medlin, L. K., & Mann, D. G. (2007). The origin and evolution of the diatoms: their adaptation to a planktonic existence. Evolution of primary producers in the sea, 207-249.

Kooistra, W. H., Sarno, D., Balzano, S., Gu, H., Andersen, R. A., & Zingone, A. (2008). Global diversity and biogeography of Skeletonema species (Bacillariophyta). Protist, 159(2), 177-193.

Korhonen, M., Rudels, B., Marnela, M., Wisotzki, A., & Zhao, J. (2013). Time and space variability of freshwater content, heat content and seasonal ice melt in the Arctic Ocean from 1991 to 2011. Ocean Science, 9(6), 1015-1055.

Krause, J. W., & Lomas, M. W. (2020). Understanding Diatoms' Past and Future Biogeochemical Role in High-Latitude Seas. Geophysical Research Letters, 47(1), e2019GL085602.

Krisch, S., Browning, T. J., Graeve, M., Ludwichowski, K. U., Lodeiro, P., Hopwood, M. J., ... & Achterberg, E. P. (2020). The influence of Arctic Fe and Atlantic fixed N on summertime primary production in Fram Strait, North Greenland Sea. Scientific reports, 10(1), 1-13.

Kröger, N., & Poulsen, N. (2008). Diatoms—from cell wall biogenesis to nanotechnology. Annual review of genetics, 42, 83-107.

Krug, L., Erlacher, A., Markut, K., Berg, G., & Cernava, T. (2020). The microbiome of alpine snow algae shows a specific inter-kingdom connectivity and algae-bacteria interactions with supportive capacities. The ISME journal, 14(9), 2197-2210.

Kumar, M. S., Slud, E. V., Okrah, K., Hicks, S. C., Hannenhalli, S., & Corrada Bravo, H. (2018). Analysis and correction of compositional bias in sparse sequencing count data. BMC genomics, 19(1), 1-23.

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. Molecular biology and evolution, 35(6), 1547.

Kustka, A. B., Allen, A. E., & Morel, F. M. (2007). Sequence analysis and transcriptional regulation of iron acquisition genes in two marine diatoms 1. Journal of Phycology, 43(4), 715-729.

Kwok, R., Zwally, H. J., & Yi, D. (2004). ICESat observations of Arctic sea ice: A first look. Geophysical Research Letters, 31(16).

Lafond, A., Leblanc, K., Quéguiner, B., Moriceau, B., Leynaert, A., Cornet, V., ... & Michel, C. (2019). Late spring bloom development of pelagic diatoms in Baffin Bay. Elementa: Science of the Anthropocene, 7.

Laland, K., Uller, T., Feldman, M., Sterelny, K., Müller, G. B., Moczek, A., ... & Strassmann, J. E. (2014). Does evolutionary theory need a rethink?. Nature News, 514(7521), 161.

Lamari, N., Ruggiero, M. V., d'Ippolito, G., Kooistra, W. H., Fontana, A., & Montresor, M. (2013). Specificity of lipoxygenase pathways supports species delineation in the marine diatom genus Pseudo-nitzschia. PLoS One, 8(8), e73281.

Lannuzel, D., Vancoppenolle, M., Van der Merwe, P., De Jong, J., Meiners, K. M., Grotti, M., ... & Miller, L. A. (2016). Iron in sea ice: Review and new insightsIron in sea ice: Review and new insights. Elementa: Science of the Anthropocene, 4.

Laso-Jadart, R., Ambroise, C., Peterlongo, P., & Madoui, M. A. (2020). metaVaR: introducing metavariant species models for reference-free metagenomic-based population genomics. PloS one, 15(12), e0244637.

Lazzardi, S., Valle, F., Mazzolini, A., Scialdone, A., Caselle, M., & Osella, M. (2021). Emergent Statistical Laws in Single-Cell Transcriptomic Data. bioRxiv.

Le Bescot, N., Mahé, F., Audic, S., Dimier, C., Garet, M. J., Poulain, J., ... & Siano, R. (2016). Global patterns of pelagic dinoflagellate diversity across protist size classes unveiled by metabarcoding. Environmental Microbiology, 18(2), 609-626.

Leblanc, K., Queguiner, B., Diaz, F., Cornet, V., Michel-Rodriguez, M., de Madron, X. D., ... & Conan, P. (2018). Nanoplanktonic diatoms are globally overlooked but play a role in spring blooms and carbon export. Nature communications, 9(1), 1-12.

Leconte, J., Benites, L. F., Vannier, T., Wincker, P., Piganeau, G., & Jaillon, O. (2020). Genome resolved biogeography of mamiellales. Genes, 11(1), 66.

Lefebvre, K. A., Quakenbush, L., Frame, E., Huntington, K. B., Sheffield, G., Stimmelmayr, R., ... & Gill, V. (2016). Prevalence of algal toxins in Alaskan marine mammals foraging in a changing arctic and subarctic environment. Harmful Algae, 55, 13-24.

Legendre, P., & Legendre, L. (2012). Numerical ecology. Elsevier.

Lelong, A., Hégaret, H., & Soudant, P. (2014). Link between domoic acid production and cell physiology after exchange of bacterial communities between toxic Pseudo-nitzschia multiseries and non-toxic Pseudo-nitzschia delicatissima. Marine drugs, 12(6), 3587-3607.

Lelong, A., Hégaret, H., Soudant, P., & Bates, S. S. (2012). Pseudo-nitzschia (Bacillariophyceae) species, domoic acid and amnesic shellfish poisoning: revisiting previous paradigms. Phycologia, 51(2), 168-216.

Lennon, J. T., & Locey, K. J. (2016). Macroecology for microbiology (No. e2478v1). PeerJ Preprints.

Liefer, J. D., Robertson, A., MacIntyre, H. L., Smith, W. L., & Dorsey, C. P. (2013). Characterization of a toxic Pseudo-nitzschia spp. bloom in the Northern Gulf of Mexico associated with domoic acid accumulation in fish. Harmful Algae, 26, 20-32.

Lim, H. C., Lim, P. T., Teng, S. T., Bates, S. S., & Leaw, C. P. (2014). Genetic structure of Pseudo-nitzschia pungens (Bacillariophyceae) populations: implications of a global diversification of the diatom. Harmful Algae, 37, 142-152.

Lim, H. C., Tan, S. N., Teng, S. T., Lundholm, N., Orive, E., David, H., ... & Leaw, C. P. (2018). Phylogeny and species delineation in the marine diatom Pseudo-nitzschia (Bacillariophyta) using cox1, LSU, and ITS 2 rRNA genes: A perspective in character evolution. Journal of phycology, 54(2), 234-248.

Lim, H. C., Teng, S. T., Lim, P. T., Wolf, M., & Leaw, C. P. (2016). 18S rDNA phylogeny of Pseudo-nitzschia (Bacillariophyceae) inferred from sequence-structure information. Phycologia, 55(2), 134-146.

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., ... & Raes, J. (2015). Determinants of community structure in the global plankton interactome. Science, 348(6237), 1262073.

Lind, S., Ingvaldsen, R. B., & Furevik, T. (2018). Arctic warming hotspot in the northern Barents Sea linked to declining sea-ice import. Nature climate change, 8(7), 634-639.

Lis, H., & Sharon, N. (1998). Lectins: carbohydrate-specific proteins that mediate cellular recognition. Chemical reviews, 98(2), 637-674.

Litchman, E., & Klausmeier, C. A. (2008). Trait-based community ecology of phytoplankton. Annual review of ecology, evolution, and systematics, 39, 615-639.

Litchman, E., de Tezanos Pinto, P., Edwards, K. F., Klausmeier, C. A., Kremer, C. T., & Thomas, M. K. (2015). Global biogeochemical impacts of phytoplankton: a trait-based perspective. Journal of ecology, 103(6), 1384-1396.

Litchman, E., Klausmeier, C. A., & Yoshiyama, K. (2009). Contrasting size evolution in marine and freshwater diatoms. Proceedings of the National Academy of Sciences, 106(8), 2665-2670.

Litchman, E., Ohman, M. D., & Kiørboe, T. (2013). Trait-based approaches to zooplankton communities. Journal of plankton research, 35(3), 473-484.

Liu, R., Wang, L., Liu, Q., Wang, Z., Li, Z., Fang, J., ... & Luo, M. (2018). Depth-resolved distribution of particle-attached and free-living bacterial communities in the water column of the New Britain Trench. Frontiers in microbiology, 9, 625.

Liu, T., Zhao, H., & Wang, T. (2020). An empirical Bayes approach to normalization and differential abundance testing for microbiome data. BMC bioinformatics, 21(1), 1-18.

Locey, K. J., & Lennon, J. T. (2016). Scaling laws predict global microbial diversity. Proceedings of the National Academy of Sciences, 113(21), 5970-5975.

Loeffler, C., Karlsberg, A., Martin, L. S., Eskin, E., Koslicki, D., & Mangul, S. (2020). Improving the usability and comprehensiveness of microbial databases. BMC biology, 18(1), 1-6.

Logares, R., Mangot, J. F., & Massana, R. (2015). Rarity in aquatic microbes: placing protists on the map. Research in microbiology, 166(10), 831-841.

Lommer, M., Specht, M., Roy, A. S., Kraemer, L., Andreson, R., Gutowska, M. A., ... & LaRoche, J. (2012). Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. Genome biology, 13(7), 1-21.

Loreau, M., Naeem, S., Inchausti, P., Bengtsson, J., Grime, J. P., Hector, A., ... & Wardle, D. A. (2001). Biodiversity and ecosystem functioning: current knowledge and future challenges. science, 294(5543), 804-808.

Louca, S., Parfrey, L. W., & Doebeli, M. (2016). Decoupling function and taxonomy in the global ocean microbiome. Science, 353(6305), 1272-1277.

Lovejoy, C. (2014). Changing views of Arctic protists (marine microbial eukaryotes) in a changing Arctic. Acta Protozoologica, 53(1).

Lovejoy, C., Legendre, L., Martineau, M. J., Bâcle, J., & Von Quillfeldt, C. H. (2002). Distribution of phytoplankton and other protists in the North Water. Deep Sea Research Part II: Topical Studies in Oceanography, 49(22-23), 5027-5047.

Lundholm, N., Bates, S. S., Baugh, K. A., Bill, B. D., Connell, L. B., Léger, C., & Trainer, V. L. (2012). Cryptic and pseudo-cryptic diversity in diatoms—with descriptions of pseudo-nitzschia hasleana sp. nov. and p. fryxelliana sp. nov. 1. Journal of Phycology, 48(2), 436-454.

Lundholm, N., Daugbjerg, N., & Moestrup, Ø. (2002). Phylogeny of the Bacillariaceae with emphasis on the genus Pseudo-nitzschia (Bacillariophyceae) based on partial LSU rDNA. European Journal of Phycology, 37(1), 115-134.

Lynch, M. D., & Neufeld, J. D. (2015). Ecology and exploration of the rare biosphere. Nature Reviews Microbiology, 13(4), 217-229.

Ma, B., Wang, Y., Ye, S., Liu, S., Stirling, E., Gilbert, J. A., ... & Xu, J. (2020). Earth microbial co-occurrence network reveals interconnection pattern across microbiomes. Microbiome, 8(1), 1-12.

Mac Arthur, R. H., & Wilson, E. O. (1967). The theory of Island biogeography (No. 574.91 M3).

Macdonald, J. D. (1869). I.—On the structure of the Diatomaceous frustule, and its genetic cycle. Journal of Natural History, 3(13), 1-8.

Madoui, M. A., Poulain, J., Sugier, K., Wessner, M., Noel, B., Berline, L., ... & Wincker, P. (2017). New insights into global biogeography, population structure and natural selection from the genome of the epipelagic copepod Oithona. Molecular ecology, 26(17), 4467-4482.

Magurran, A. E. (2013). Measuring biological diversity. John Wiley & Sons.

Magurran, A.E. (2004). Measuring biological diversity, Blackwell Science, Oxford.

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.

Maldonado, M. T., Allen, A. E., Chong, J. S., Lin, K., Leus, D., Karpenko, N., & Harris, S. L. (2006). Copper-dependent iron transport in coastal and oceanic diatoms. Limnology and oceanography, 51(4), 1729-1743.

Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., ... & Bowler, C. (2016). Insights into global diatom distribution and diversity in the world's ocean. Proceedings of the National Academy of Sciences, 113(11), E1516-E1525.

Manhart, J. R., Fryxell, G. A., Villac, M. C., & Segura, L. Y. (1995). Pseudo-nitzschia pungens and P. multisenes (Bacillariophyceae): Nuclear Ribosomal DNAs and species differences. Journal of Phycology, 31(3), 421-427.

Mann, D. G., & Round, F. (1988). Why didn't Lund see sex in Asterionella? A discussion of the diatom life cycle in nature. Algae and the aquatic environment, 29, 385-412.

Mann, D. G., & Vanormelingen, P. (2013). An inordinate fondness? The number, distributions, and origins of diatom species. Journal of eukaryotic microbiology, 60(4), 414-420.

Marchetti, A., Moreno, C. M., Cohen, N. R., Oleinikov, I., deLong, K., Twining, B. S., ... & Lampe, R. H. (2017). Development of a molecular-based index for assessing iron status in bloom-forming pennate diatoms. Journal of phycology, 53(4), 820-832.

Marchetti, A., Parker, M. S., Moccia, L. P., Lin, E. O., Arrieta, A. L., Ribalet, F., ... & Armbrust, E. (2009). Ferritin is used for iron storage in bloom-forming marine pennate diatoms. Nature, 457(7228), 467-470.

Margalef, R. (1978). Life-forms of phytoplankton as survival alternatives in an unstable environment. Oceanologica acta, 1(4), 493-509.

Marquet, P. A., Abades, S. R., & Labra, F. A. (2007). Biodiversity power laws. Scaling biodiversity, 441-461.

Martino, C., Morton, J. T., Marotz, C. A., Thompson, L. R., Tripathi, A., Knight, R., & Zengler, K. (2019). A novel sparse compositional technique reveals microbial perturbations. MSystems, 4(1), e00016-19.

May, F., Huth, A., & Wiegand, T. (2015). Moving beyond abundance distributions: neutral theory and spatial patterns in a tropical forest. Proceedings of the Royal Society B: Biological Sciences, 282(1802), 20141657.

McCabe, R. M., Hickey, B. M., Kudela, R. M., Lefebvre, K. A., Adams, N. G., Bill, B. D., ... & Trainer, V. L. (2016). An unprecedented coastwide toxic algal bloom linked to anomalous ocean conditions. Geophysical Research Letters, 43(19), 10-366.

McCann, K. S. (2000). The diversity–stability debate. Nature, 405(6783), 228-233.

McClelland, J. W., Holmes, R. M., Dunton, K. H., & Macdonald, R. W. (2012). The Arctic ocean estuary. Estuaries and Coasts, 35(2), 353-368.

McDonald, S. M., Sarno, D., & Zingone, A. (2007). Identifying Pseudo-nitzschia species in natural samples using genus-specific PCR primers and clone libraries. Harmful Algae, 6(6), 849-860.

McGill, B. J. (2019). The what, how and why of doing macroecology. Global Ecology and Biogeography, 28(1), 6-17.

McGill, B. J., Etienne, R. S., Gray, J. S., Alonso, D., Anderson, M. J., Benecha, H. K., ... & White, E. P. (2007). Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. Ecology letters, 10(10), 995-1015.

McNaughton, S. J., & Wolf, L. L. (1970). Dominance and the niche in ecological systems. Science, 167(3915), 131-139.

Menon, S., Denman, K. L., Brasseur, G., Chidthaisong, A., Ciais, P., Cox, P. M., ... & Zhang, X. (2007). Couplings between changes in the climate system and biogeochemistry (No. LBNL-464E). Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).

Meredith, M., Sommerkorn, M., Cassotta, S., Derksen, C., Ekaykin, A., Hollowed, A., ... & Schuur, E. A. (2019). Polar Regions. Chapter 3, IPCC Special Report on the Ocean and Cryosphere in a Changing Cli-mate.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2020). Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071). TU Wien. R package version 1.7-2.: e1071.

Meyer, K. M., Memiaghe, H., Korte, L., Kenfack, D., Alonso, A., & Bohannan, B. J. (2018). Why do microbes exhibit weak biogeographic patterns?. The ISME journal, 12(6), 1404-1413.

Mikhailov, I. S., Zakharova, Y. R., Bukin, Y. S., Galachyants, Y. P., Petrova, D. P., Sakirko, M. V., & Likhoshway, Y. V. (2019). Co-occurrence networks among bacteria and microbial eukaryotes of Lake Baikal during a spring phytoplankton bloom. Microbial ecology, 77(1), 96-109.

MillerO, F. J., Woosley, R., DiTrolio, B., & Waters, J. (2009). Effect of ocean acidification on the speciation of metals in seawater. Oceanography, 22(4), 72-85.

Mock, T., Krell, A., Glöckner, G., Kolukisaoglu, Ü., & Valentin, K. (2006). Analysis Of Expressed Sequence Tags (Ests) From The Polar Diatom Fragilariopsis cylindrus 1. Journal of Phycology, 42(1), 78-85.

Mock, T., Otillar, R. P., Strauss, J., McMullan, M., Paajanen, P., Schmutz, J., ... & Grigoriev, I. V. (2017). Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus. Nature, 541(7638), 536-540.

Mohit, V., Archambault, P., Toupoint, N., & Lovejoy, C. (2014). Phylogenetic differences in attached and free-living bacterial communities in a temperate coastal lagoon during summer, revealed via high-throughput 16S rRNA gene sequencing. Applied and environmental microbiology, 80(7), 2071-2083.

Monchamp, M. E., Spaak, P., & Pomati, F. (2018). High dispersal levels and lake warming are emergent drivers of cyanobacterial community assembly over the Anthropocene in peri-Alpine lakes. bioRxiv, 419762.

Monier, A., Worden, A. Z., & Richards, T. A. (2016). Phylogenetic diversity and biogeography of the Mamiellophyceae lineage of eukaryotic phytoplankton across the oceans. Environmental microbiology reports, 8(4), 461-469.

Montresor, M., Vitale, L., D'Alelio, D., & Ferrante, M. I. (2016). Sex in marine planktonic diatoms: insights and challenges. Perspect. Phycol, 3(2), 61-75.

Moore, C. M., Mills, M. M., Achterberg, E. P., Geider, R. J., LaRoche, J., Lucas, M. I., ... & Woodward, E. M. S. (2009). Large-scale distribution of Atlantic nitrogen fixation controlled by iron availability. Nature Geoscience, 2(12), 867-871.

Moore, J. K., & Doney, S. C. (2007). Iron availability limits the ocean nitrogen inventory stabilizing feedbacks between marine denitrification and nitrogen fixation. Global Biogeochemical Cycles, 21(2).

Moore, J. K., Doney, S. C., & Lindsay, K. (2004). Upper ocean ecosystem dynamics and iron cycling in a global three-dimensional model. Global Biogeochemical Cycles, 18(4).

Moore, J. K., Doney, S. C., Glover, D. M., & Fung, I. Y. (2001). Iron cycling and nutrient-limitation patterns in surface waters of the World Ocean. Deep Sea Research Part II: Topical Studies in Oceanography, 49(1-3), 463-507.

Mordret, S., Piredda, R., Vaulot, D., Montresor, M., Kooistra, W. H., & Sarno, D. (2018). dinoref: A curated dinoflagellate (Dinophyceae) reference database for the 18S rRNA gene. Molecular Ecology Resources, 18(5), 974-987.

Morel, F. M., & Price, N. M. (2003). The biogeochemical cycles of trace metals in the oceans. Science, 300(5621), 944-947.

Moreno, C. M., Lin, Y., Davies, S., Monbureau, E., Cassar, N., & Marchetti, A. (2018). Examination of gene repertoires and physiological responses to iron and light limitation in Southern Ocean diatoms. Polar Biology, 41(4), 679-696.

Morris, E. K., Caruso, T., Buscot, F., Fischer, M., Hancock, C., Maier, T. S., ... & Socher, S. A. (2014). Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories. Ecology and evolution, 4(18), 3514-3524.

Moustafa, A., Beszteri, B., Maier, U. G., Bowler, C., Valentin, K., & Bhattacharya, D. (2009). Genomic footprints of a cryptic plastid endosymbiosis in diatoms. science, 324(5935), 1724-1726.

Muilwijk, M., Smedsrud, L. H., Ilicak, M., & Drange, H. (2018). Atlantic Water heat transport variability in the 20th century Arctic Ocean from a global ocean model and observations. Journal of Geophysical Research: Oceans, 123(11), 8159-8179.

Mukhopadhyay, S., & Jackson, P. K. (2011). The tubby family proteins. Genome biology, 12(6), 1-9.

Nakov, T., Beaulieu, J. M., & Alverson, A. J. (2018). Insights into global planktonic diatom diversity: The importance of comparisons between phylogenetically equivalent units that account for time. The ISME journal, 12(11), 2807-2810.

Nanjappa, D., Audic, S., Romac, S., Kooistra, W. H., & Zingone, A. (2014). Assessment of species diversity and distribution of an ancient diatom lineage using a DNA metabarcoding approach. PLoS One, 9(8), e103810.

Nanjappa, D., Kooistra, W. H., & Zingone, A. (2013). A reappraisal of the genus Leptocylindrus (Bacillariophyta), with the addition of three species and the erection of T enuicylindrus gen. nov. Journal of phycology, 49(5), 917-936.

Nedelcu, A. M., Marcu, O., & Michod, R. E. (2004). Sex as a response to oxidative stress: a twofold increase in cellular reactive oxygen species activates sex genes. Proceedings of the Royal Society of London. Series B: Biological Sciences, 271(1548), 1591-1596.

Needham, D. M., Chow, C. E. T., Cram, J. A., Sachdeva, R., Parada, A., & Fuhrman, J. A. (2013). Short-term observations of marine bacterial and viral communities: patterns, connections and resilience. The ISME journal, 7(7), 1274-1285.

Nelson, D. M., Tréguer, P., Brzezinski, M. A., Leynaert, A., & Quéguiner, B. (1995). Production and dissolution of biogenic silica in the ocean: revised global estimates, comparison with regional data and relationship to biogenic sedimentation. Global biogeochemical cycles, 9(3), 359-372.

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution, 32(1), 268-274.

Nielsdóttir, M. C., Moore, C. M., Sanders, R., Hinz, D. J., & Achterberg, E. P. (2009). Iron limitation of the postbloom phytoplankton communities in the Iceland Basin. Global Biogeochemical Cycles, 23(3).

Nisbet, R. E. R., Kilian, O., & McFadden, G. I. (2004). Diatom genomics: genetic acquisitions and mergers. Current Biology, 14(24), R1048-R1050.

Nuester, J., Vogt, S., & Twining, B. S. (2012). Localization of iron within centric diatoms of the genus Thalassiosira. Journal of phycology, 48(3), 626-634.

Ogura, A., Akizuki, Y., Imoda, H., Mineta, K., Gojobori, T., & Nagai, S. (2018). Comparative genome and transcriptome analysis of diatom, Skeletonema costatum, reveals evolution of genes for harmful algal bloom. BMC genomics, 19(1), 1-12.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'hara, R. B., ... & Oksanen, M. J. (2013). Package 'vegan'. Community ecology package, version, 2(9), 1-295.

Osuna-Cruz, C. M., Bilcke, G., Vancaester, E., De Decker, S., Bones, A. M., Winge, P., ... & Vandepoele, K. (2020). The Seminavis robusta genome provides insights into the evolutionary adaptations of benthic diatoms. Nature communications, 11(1), 1-13.

Oziel, L., Baudena, A., Ardyna, M., Massicotte, P., Randelhoff, A., Sallée, J. B., ... & Babin, M. (2020). Faster Atlantic currents drive poleward expansion of temperate phytoplankton in the Arctic Ocean. Nature communications, 11(1), 1-8.

Pallmann, P., Schaarschmidt, F., Hothorn, L. A., Fischer, C., Nacke, H., Priesnitz, K. U., & Schork, N. J. (2012). Assessing group differences in biodiversity by simultaneously testing a user-defined selection of diversity indices. Molecular ecology resources, 12(6), 1068-1078.

Paquette, R. G., Bourke, R. H., Newton, J. F., & Perdue, W. F. (1985). The East Greenland polar front in autumn. Journal of Geophysical Research: Oceans, 90(C3), 4866-4882.

Parks, M. B., Nakov, T., Ruck, E. C., Wickett, N. J., & Alverson, A. J. (2018). Phylogenomics reveals an extensive history of genome duplication in diatoms (Bacillariophyta). American Journal of Botany, 105(3), 330-347.

Passy, S. I. (2016). Abundance inequality in freshwater communities has an ecological origin. The American Naturalist, 187(4), 502-516.

Patil, S., Moeys, S., von Dassow, P., Huysman, M. J., Mapleson, D., De Veylder, L., ... & Ferrante, M. I. (2015). Identification of the meiotic toolkit in diatoms and exploration of meiosis-specific SPO11 and RAD51 homologs in the sexual species Pseudo-nitzschia multistriata and Seminavis robusta. BMC genomics, 16(1), 1-21.

Paul, C., & Pohnert, G. (2011). Interactions of the algicidal bacterium Kordia algicida with diatoms: regulated protease excretion for specific algal lysis. PloS one, 6(6), e21032.

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., ... & De Vargas, C. (2012). CBOL protist working group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. PLoS biology, 10(11), e1001419.

Pearman, J. K., Casas, L., Merle, T., Michell, C., & Irigoien, X. (2016). Bacterial and protist community changes during a phytoplankton bloom. Limnology and Oceanography, 61(1), 198-213.

Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. Current protocols in bioinformatics, 42(1), 3-1.

Pedrós-Alió, C. (2006). Marine microbial diversity: can it be determined?. Trends in microbiology, 14(6), 257-263.

Percopo, I., Ruggiero, M. V., Balzano, S., Gourvil, P., Lundholm, N., Siano, R., ... & Sarno, D. (2016). Pseudo-nitzschia arctica sp. nov., a new cold-water cryptic Pseudo-nitzschia species within the P. pseudodelicatissima complex. Journal of Phycology, 52(2), 184-199.

Percopo, I., Ruggiero, M. V., Sarno, D., Longobardi, L., Rossi, R., Piredda, R., & Zingone, A. (2021). Phenological segregation suggests speciation by time in the planktonic diatom Pseudo-nitzschia allochrona sp. nov. bioRxiv.

Perez-Riba, A., & Itzhaki, L. S. (2019). The tetratricopeptide-repeat motif is a versatile platform that enables diverse modes of molecular recognition. Current opinion in structural biology, 54, 43-49.

Perovich, D. K., & Polashenski, C. (2012). Albedo evolution of seasonal Arctic sea ice. Geophysical Research Letters, 39(8).

Perrette, M., Yool, A., Quartly, G. D., & Popova, E. E. (2011). Near-ubiquity of ice-edge blooms in the Arctic. Biogeosciences, 8(2), 515-524.

Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., ... & Searson, S. (2015). Open science resources for the discovery and analysis of Tara Oceans data. Scientific data, 2(1), 1-16.

Pfaffen, S., Abdulqadir, R., Le Brun, N. E., & Murphy, M. E. (2013). Mechanism of ferrous iron binding and oxidation by ferritin from a pennate diatom. Journal of Biological Chemistry, 288(21), 14917-14925.

Pfitzer, E. (1869). Über den Bau und die Zellteilung der Diatomeen. Botanische Zeitung, 27(46), 774-776.

Pierson, K. (2020). Building a richer understanding of diversity through causally consistent evenness measures. Ecology and Evolution.

Piredda, R., Claverie, J. M., Decelle, J., de Vargas, C., Dunthorn, M., Edvardsen, B., ... & Zingone, A. (2018). Diatom diversity through HTS-metabarcoding in coastal European seas. Scientific Reports, 8(1), 1-12.

Piredda, R., Tomasino, M. P., D'erchia, A. M., Manzari, C., Pesole, G., Montresor, M., ... & Zingone, A. (2017). Diversity and temporal patterns of planktonic protist assemblages at a Mediterranean Long Term Ecological Research site. FEMS microbiology ecology, 93(1).

Planes, S., Allemand, D., Agostini, S., Banaigs, B., Boissin, E., Boss, E., ... & Tara Pacific Consortium. (2019). The Tara Pacific expedition—A pan-ecosystemic approach of the "-omics" complexity of coral reef holobionts across the Pacific Ocean. PLoS biology, 17(9), e3000483.

Podani, J., & Schmera, D. (2011). A new conceptual and methodological framework for exploring and explaining pattern in presence–absence data. Oikos, 120(11), 1625-1638.

Podani, J., Ódor, P., Fattorini, S., Strona, G., Heino, J., & Schmera, D. (2018). Exploring multiple presence-absence data structures in ecology. Ecological Modelling, 383, 41-51.

Poloczanska, E. S., Brown, C. J., Sydeman, W. J., Kiessling, W., Schoeman, D. S., Moore, P. J., ... & Richardson, A. J. (2013). Global imprint of climate change on marine life. Nature Climate Change, 3(10), 919-925.

Polyak, L., Alley, R. B., Andrews, J. T., Brigham-Grette, J., Cronin, T. M., Darby, D. A., ... & Wolff, E. (2010). History of sea ice in the Arctic. Quaternary Science Reviews, 29(15-16), 1757-1778.

Polyakov, I. V., Pnyushkov, A. V., Alkire, M. B., Ashik, I. M., Baumann, T. M., Carmack, E. C., ... & Yulin, A. (2017). Greater role for Atlantic inflows on sea-ice loss in the Eurasian Basin of the Arctic Ocean. Science, 356(6335), 285-291.

Pondaven, P., Ragueneau, O., Tréguer, P., Hauvespre, A., Dezileau, L., & Reyss, J. L. (2000). Resolving the 'opal paradox'in the Southern Ocean. Nature, 405(6783), 168-172.

Poulsen, N. C., Spector, I., Spurck, T. P., Schultz, T. F., & Wetherbee, R. (1999). Diatom gliding is the result of an actin-myosin motility system. Cell motility and the cytoskeleton, 44(1), 23-33.

Preston, F. W. (1948). The commonness, and rarity, of species. Ecology, 29(3), 254-283.

Preston, F. W. (1962). The canonical distribution of commonness and rarity: Part I. Ecology, 43(2), 185-215.

Prowe, A. F., Pahlow, M., Dutkiewicz, S., Follows, M., & Oschlies, A. (2012). Top-down control of marine phytoplankton diversity in a global ecosystem model. Progress in Oceanography, 101(1), 1-13.

Pugliese, L., Casabianca, S., Perini, F., Andreoni, F., & Penna, A. (2017). A high resolution melting method for the molecular identification of the potentially toxic diatom Pseudo-nitzschia spp. in the Mediterranean Sea. Scientific Reports, 7(1), 1-10.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., ... & Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research, 41(D1), D590-D596.

Quéguiner, B. Iron fertilization and the structure of planktonic communities in high nutrient regions of the Southern Ocean. Deep-Sea Res. II 90, 43–54 (2013).

Quéré, C. L., Harrison, S. P., Colin Prentice, I., Buitenhuis, E. T., Aumont, O., Bopp, L., ... & Wolf-Gladrow, D. (2005). Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. Global Change Biology, 11(11), 2016-2040.

Quijano-Scheggia, S. I., Garcés, E., Lundholm, N., Moestrup, Ø., Andree, K., & Camp, J. (2009). Morphology, physiology, molecular phylogeny and sexual compatibility of the cryptic Pseudo-nitzschia delicatissima complex (Bacillariophyta), including the description of P. arenysensis sp. nov. Phycologia, 48*(6), 492-509.*

Rabosky, D. L., Chang, J., Title, P. O., Cowman, P. F., Sallan, L., Friedman, M., ... & Alfaro, M. E. (2018). An inverse latitudinal gradient in speciation rate for marine fishes. Nature, 559(7714), 392-395.

Ragueneau, O., Tréguer, P., Leynaert, A., Anderson, R. F., Brzezinski, M. A., DeMaster, D. J., ... & Quéguiner, B. (2000). A review of the Si cycle in the modern ocean: recent progress and missing gaps in the application of biogenic opal as a paleoproductivity proxy. Global and Planetary Change, 26(4), 317-365.

Rambaut A, Drummond, A (2012) Fig Tree version 1.4.0. htt://tree.bio.ed.ac.uk/software/

Ramond, P., Sourisseau, M., Simon, N., Romac, S., Schmitt, S., Rigaut-Jalabert, F., ... & Siano, R. (2019). Coupling between taxonomic and functional diversity in protistan coastal communities. Environmental microbiology, 21(2), 730-749.

Ramsayer, J., Fellous, S., Cohen, J. E., & Hochberg, M. E. (2012). Taylor's law holds in experimental bacterial populations but competition does not influence the slope. Biology letters, 8(2), 316-319.

Randelhoff, A., Holding, J., Janout, M., Sejr, M. K., Babin, M., Tremblay, J. É., & Alkire, M. B. (2020). Pan-Arctic Ocean primary production constrained by turbulent nitrate fluxes. Frontiers in Marine Science, 7, 150.

Rengefors, K., Kremp, A., Reusch, T. B., & Wood, A. M. (2017). Genetic diversity and evolution in eukaryotic phytoplankton: revelations from population genetic studies. Journal of Plankton Research, 39(2), 165-179.

Richerson, P., Armstrong, R., & Goldman, C. R. (1970). Contemporaneous disequilibrium, a new hypothesis to explain the „Paradox of the Plankton". Proceedings of the National Academy of Sciences, 67(4), 1710-1714.

Rieck, A., Herlemann, D. P., Jürgens, K., & Grossart, H. P. (2015). Particle-associated differ from free-living bacteria in surface waters of the Baltic Sea. Frontiers in Microbiology, 6, 1297.

Rijkenberg, M. J., Middag, R., Laan, P., Gerringa, L. J., van Aken, H. M., Schoemann, V., ... & De Baar, H. J. (2014). The distribution of dissolved iron in the West Atlantic Ocean. PloS one, 9(6), e101323.

Rimet, F., Gusev, E., Kahlert, M., Kelly, M. G., Kulikovskiy, M., Maltsev, Y., ... & Bouchez, A. (2019). Diat. barcode, an open-access curated barcode library for diatoms. Scientific Reports, 9(1), 1-12.

Robinson, J. V., & Sandgren, C. D. (1983). The effect of temporal environmental heterogeneity on community structure: a replicated experimental study. Oecologia, 57(1), 98-102.

Round, F. E., Crawford, R. M., & Mann, D. G. (1990). Diatoms: biology and morphology of the genera. Cambridge university press.

Royo-Llonch, Marta, Pablo Sánchez, Clara Ruiz-González, Guillem Salazar, Carlos Pedrós-Alió, Karine Labadie, Lucas Paoli et al. "Ecogenomics of key prokaryotes in the arctic ocean." (2020).

Rudels, B., Anderson, L. G., & Jones, E. P. (1996). Formation and evolution of the surface mixed layer and halocline of the Arctic Ocean. Journal of Geophysical Research: Oceans, 101(C4), 8807-8821.

Rudels, B., Schauer, U., Björk, G., Korhonen, M., Pisarev, S., Rabe, B., & Wisotzki, A. (2013). Observations of water masses and circulation with focus on the Eurasian Basin of the Arctic Ocean from the 1990s to the late 2000s. Ocean Science, 9(1), 147-169.

Ruggiero, M. V., Kooistra, W. H. C. F., Piredda, R., Sarno, D., Zampicinini, G., Zingone, A., & Montresor, M. (2022). Temporal changes of genetic structure and diversity in a marine diatom genus discovered via metabarcoding. Environmental DNA, 00, 1– 13.

Ruggiero, M. V., Sarno, D., Barra, L., Kooistra, W. H., Montresor, M., & Zingone, A. (2015). Diversity and temporal pattern of Pseudo-nitzschia species (Bacillariophyceae) through the molecular lens. Harmful Algae, 42, 15-24.

Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., ... & Venter, J. C. (2007). The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS biology, 5(3), e77.

Russo, M. T., Vitale, L., Entrambasaguas, L., Anestis, K., Fattorini, N., Romano, F., ... & Ferrante, M. I. (2018). MRP3 is a sex determining gene in the diatom Pseudo-nitzschia multistriata. Nature communications, 9(1), 1-10.

Ryan, J. P., Kudela, R. M., Birch, J. M., Blum, M., Bowers, H. A., Chavez, F. P., ... & Zhang, Y. (2017). Causality of an extreme harmful algal bloom in Monterey Bay, California, during the 2014–2016 northeast Pacific warm anomaly. Geophysical Research Letters, 44(11), 5571-5579.

Ryan-Keogh, T. J., Macey, A. I., Nielsdóttir, M. C., Lucas, M. I., Steigenberger, S. S., Stinchcombe, M. C., ... & Moore, C. M. (2013). Spatial and temporal development of phytoplankton iron stress in relation to bloom dynamics in the high-latitude North Atlantic Ocean. Limnology and oceanography, 58(2), 533-545.

Rynearson, T. A., & Armbrust, E. V. (2000). DNA fingerprinting reveals extensive genetic diversity in a field population of the centric diatom Ditylum brightwellii. Limnology and Oceanography, 45(6), 1329-1340.

Rynearson, T. A., Newton, J. A., & Armbrust, E. V. (2006). Spring bloom development, genetic variation, and population succession in the planktonic diatom Ditylum brightwellii. Limnology and Oceanography, 51(3), 1249-1261.

Sakshaug, E. (2004). Primary and secondary production in the Arctic Seas. In The organic carbon cycle in the Arctic Ocean (pp. 57-81). Springer, Berlin, Heidelberg.

Salazar, G., Paoli, L., Alberti, A., Huerta-Cepas, J., Ruscheweyh, H. J., Cuenca, M., ... & Wincker, P. (2019). Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. Cell, 179(5), 1068-1083.

Santoferrara, L., Burki, F., Filker, S., Logares, R., Dunthorn, M., & McManus, G. B. (2020). Perspectives from ten years of protist studies by high-throughput metabarcoding. Journal of Eukaryotic Microbiology, 67(5), 612-622.

Sarno, D., Kooistra, W. H., Balzano, S., Hargraves, P. E., & Zingone, A. (2007). Diversity in the genus Skeletonema (BACILLARIOPHYCEAE): III. Phylogenetic position and morphological variability of Skeletonema costatum and Skeletonema grevillei, with the description of Skeletonema ardens sp. NOV. 1. Journal of phycology, 43(1), 156-170.

Sarno, D., Kooistra, W. H., Medlin, L. K., Percopo, I., & Zingone, A. (2005). Diversity in the genus Skeletonema (Bacillariophyceae). ii. an assessment of the taxonomy of S. costatum-like species with the description of four new species 1. Journal of phycology, 41(1), 151-176.

Satinsky, B. M., Gifford, S. M., Crump, B. C., & Moran, M. A. (2013). Use of internal standards for quantitative metatranscriptome and metagenome analysis. In Methods in enzymology (Vol. 531, pp. 237-250). Academic Press.

Sato, S., Beakes, G., Idei, M., Nagumo, T., & Mann, D. G. (2011). Novel sex cells and evidence for sex pheromones in diatoms. PloS one, 6(10), e26923.

Scalco, E., Amato, A., Ferrante, M. I., & Montresor, M. (2016). The sexual phase of the diatom Pseudo-nitzschia multistriata: cytological and time-lapse cinematography characterization. Protoplasma, 253(6), 1421-1431.

Scalco, E., Stec, K., Iudicone, D., Ferrante, M. I., & Montresor, M. (2014). The dynamics of sexual phase in the marine diatom P seudo-nitzschia multistriata (B acillariophyceae). Journal of phycology, 50(5), 817-828.

Schauer, U., Loeng, H., Rudels, B., Ozhigin, V. K., & Dieck, W. (2002). Atlantic water flow through the Barents and Kara Seas. Deep Sea Research Part I: Oceanographic Research Papers, 49(12), 2281-2298.

Scheffer, M., & van Nes, E. H. (2006). Self-organized similarity, the evolutionary emergence of groups of similar species. Proceedings of the National Academy of Sciences, 103(16), 6230-6235.

Schourup-Kristensen, V., Wekerle, C., Wolf-Gladrow, D. A., & Völker, C. (2018). Arctic Ocean biogeochemistry in the high resolution FESOM 1.4-REcoM2 model. Progress in Oceanography, 168, 65-81.

Ser-Giacomi, E., Zinger, L., Malviya, S., De Vargas, C., Karsenti, E., Bowler, C., & De Monte, S. (2018). Ubiquitous abundance distribution of non-dominant plankton across the global ocean. Nature ecology & evolution, 2(8), 1243-1249.

Seymour, J. R., Amin, S. A., Raina, J. B., & Stocker, R. (2017). Zooming in on the phycosphere: the ecological interface for phytoplankton–bacteria relationships. Nature microbiology, 2(7), 1-12.

Shade, A., Dunn, R. R., Blowes, S. A., Keil, P., Bohannan, B. J., Herrmann, M., ... & Chase, J. (2018). Macroecology to unite all life, large and small. Trends in ecology & evolution, 33(10), 731-744.

Shannon, C. E. (1948). A mathematical theory of communication. The Bell system technical journal, 27(3), 379-423.

Shi, D., Xu, Y., Hopkinson, B. M., & Morel, F. M. (2010). Effect of ocean acidification on iron availability to marine phytoplankton. Science, 327(5966), 676-679.

Shim, E., Shim, J., Klochkova, T. A., Han, J. W., & Kim, G. H. (2012). Purification of a sex-specific lectin involved in gamete binding of aglaothamnion callophyllidicola (rhodophyta) 1. Journal of phycology, 48(4), 916-924.

Siegel, D. A. (1998). Resource competition in a discrete environment: Why are plankton distributions paradoxical?. Limnology and Oceanography, 43(6), 1133-1146.

Simpson, E. H. (1949). Measurement of diversity. Nature, 163(4148), 688-688.

Sison-Mangus, M. P., Jiang, S., Tran, K. N., & Kudela, R. M. (2014). Host-specific adaptation governs the interaction of the marine diatom, Pseudo-nitzschia and their microbiota. The ISME journal, 8(1), 63-76.

Sjöqvist, C., Godhe, A., Jonsson, P. R., Sundqvist, L., & Kremp, A. (2015). Local adaptation and oceanographic connectivity patterns explain genetic differentiation of a marine diatom across the North Sea–Baltic Sea salinity gradient. Molecular ecology, 24(11), 2871-2885.

Smetacek, V. (1999). Diatoms and the ocean carbon cycle. Protist, 150(1), 25-32.

Smetacek, V. (2001). A watery arms race. Nature, 411(6839), 745-745.

Smith, B., & Wilson, J. B. (1996). A consumer's guide to evenness indices. Oikos, 70-82.

Smith, S. L., Pahlow, M., Merico, A., Acevedo-Trejos, E., Sasai, Y., Yoshikawa, C., ... & Honda, M. C. (2016). Flexible phytoplankton functional type (FlexPFT) model: size-scaling of traits and optimal growth. Journal of Plankton Research, 38(4), 977-992.

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., ... & Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". Proceedings of the National Academy of Sciences, 103(32), 12115-12120.

Soininen, J., & Teittinen, A. (2019). Fifteen important questions in the spatial ecology of diatoms. Freshwater Biology, 64(11), 2071-2083.

Soininen, J., Jamoneau, A., Rosebery, J., & Passy, S. I. (2016). Global patterns and drivers of species and trait composition in diatoms. Global ecology and biogeography, 25(8), 940-950.

Sommer, U. (1984). The paradox of the plankton: Fluctuations of phosphorus availability maintain diversity of phytoplankton in flow-through cultures 1. Limnology and Oceanography, 29(3), 633-636.

Sommer, U. (1989). The role of competition for resources in phytoplankton succession. In Plankton ecology (pp. 57-106). Springer, Berlin, Heidelberg.

Sommeria-Klein, G., Watteaux, R., Ibarbalz, F. M., Pierella Karlusich, J. J., Iudicone, D., Bowler, C., & Morlon, H. (2021). Global drivers of eukaryotic plankton biogeography in the sunlit ocean. Science, 374(6567), 594-599.

Spencer, H. (1872). The Survival of the Fittest. Nature, 5(118), 263-264.

Steinacher, M., Joos, F., Frölicher, T. L., Bopp, L., Cadule, P., Cocco, V., ... & Segschneider, J. (2010). Projected 21st century decrease in marine productivity: a multi-model analysis. Biogeosciences, 7(3), 979-1005.

Stern, R., Kraberg, A., Bresnan, E., Kooistra, W. H., Lovejoy, C., Montresor, M., ... & Metfies, K. (2018). Molecular analyses of protists in long-term observation programmes—current status and future perspectives. Journal of Plankton Research, 40(5), 519-536.

Stock, C. A., Dunne, J. P., & John, J. G. (2014). Drivers of trophic amplification of ocean productivity trends in a changing climate. Biogeosciences, 11(24), 7125-7135.

Stukel, M. R., Coles, V. J., Brooks, M. T., & Hood, R. R. (2014). Top-down, bottom-up and physical controls on diatom-diazotroph assemblage growth in the Amazon River plume. Biogeosciences, 11(12), 3259-3278.

Sunagawa, S., Acinas, S. G., Bork, P., Bowler, C., Eveillard, D., Gorsky, G., ... & de Vargas, C. (2020). Tara Oceans: towards global ocean ecosystems biology. Nature Reviews Microbiology, 18(8), 428-445.

Swan, B. K., Tupper, B., Sczyrba, A., Lauro, F. M., Martinez-Garcia, M., González, J. M., ... & Stepanauskas, R. (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. Proceedings of the National Academy of Sciences, 110(28), 11463-11468.

Tagliabue, A., Bowie, A. R., Boyd, P. W., Buck, K. N., Johnson, K. S., & Saito, M. A. (2017). The integral role of iron in ocean biogeochemistry. Nature, 543(7643), 51-59.

Tammilehto, A., Nielsen, T. G., Krock, B., Møller, E. F., & Lundholm, N. (2015). Induction of domoic acid production in the toxic diatom Pseudo-nitzschia seriata by calanoid copepods. Aquatic Toxicology, 159, 52-61.

Tanaka, T., Maeda, Y., Veluchamy, A., Tanaka, M., Abida, H., Maréchal, E., ... & Fujibuchi, W. (2015). Oil accumulation by the oleaginous diatom Fistulifera solaris as revealed by the genome and transcriptome. The Plant Cell, 27(1), 162-176.

Taylor, G. T., & Sullivan, C. W. (2008). Vitamin B12 and cobalt cycling among diatoms and bacteria in Antarctic sea ice microbial communities. Limnology and Oceanography, 53(5), 1862-1877.

Taylor, L. R. (1961). Aggregation, variance and the mean. Nature, 189(4766), 732-735.

Taylor, R. L., Semeniuk, D. M., Payne, C. D., Zhou, J., Tremblay, J. É., Cullen, J. T., & Maldonado, M. T. (2013). Colimitation by light, nitrate, and iron in the Beaufort Sea in late summer. Journal of Geophysical Research: Oceans, 118(7), 3260-3277.

Team, R. C. (2019). 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria: Available at: https://www. R-project. org/.

Terrado, R., Scarcella, K., Thaler, M., Vincent, W. F., & Lovejoy, C. (2013). Small phytoplankton in Arctic seas: vulnerability to climate change. Biodiversity, 14(1), 2-18.

Terseleer, N., Bruggeman, J., Lancelot, C., & Gypens, N. (2014). Trait-based representation of diatom functional diversity in a plankton functional type model of the eutrophied southern North Sea. Limnology and Oceanography, 59(6), 1958-1972.

Tesson, S. V., Legrand, C., van Oosterhout, C., Montresor, M., Kooistra, W. H., & Procaccini, G. (2013). Mendelian inheritance pattern and high mutation rates of microsatellite alleles in the diatom Pseudo-nitzschia multistriata. Protist, 164(1), 89-100.

Thingstad, T. F. (2000). Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. Limnology and Oceanography, 45(6), 1320-1328.

Thomas, M. K., Kremer, C. T., Klausmeier, C. A., & Litchman, E. (2012). A global pattern of thermal adaptation in marine phytoplankton. Science, 338(6110), 1085-1088.

Thomsen, P. F., & Willerslev, E. (2015). Environmental DNA–An emerging tool in conservation for monitoring past and present biodiversity. Biological conservation, 183, 4-18.

Tilman, D. (1982). Resource Competition and Community Structure." Princeton Univ. Press, Princeton, New jersey.

Tilman, D., Isbell, F., & Cowles, J. M. (2014). Biodiversity and ecosystem functioning. Annual review of ecology, evolution, and systematics, 45, 471-493.

Timmermans, M. L., & Marshall, J. (2020). Understanding Arctic Ocean circulation: a review of ocean dynamics in a changing climate. Journal of Geophysical Research: Oceans, 125(4).

Timmermans, M. L., Winsor, P., & Whitehead, J. A. (2005). Deep-water flow over the Lomonosov Ridge in the Arctic Ocean. Journal of Physical Oceanography, 35(8), 1489-1493.

Tirichine, L., Rastogi, A., & Bowler, C. (2017). Recent progress in diatom genomics and epigenomics. Current opinion in plant biology, 36, 46-55.

Tragin, M., Zingone, A., & Vaulot, D. (2018). Comparison of coastal phytoplankton composition estimated from the V4 and V9 regions of the 18S rRNA gene with a focus on photosynthetic groups and especially Chlorophyta. Environmental microbiology, 20(2), 506-520.

Traller, J. C., Cokus, S. J., Lopez, D. A., Gaidarenko, O., Smith, S. R., McCrow, J. P., ... & Hildebrand, M. (2016). Genome and methylome of the oleaginous diatom Cyclotella cryptica reveal genetic flexibility toward a high lipid phenotype. Biotechnology for biofuels, 9(1), 1-20.

Trano, A. C. (2021). Exploring Marine Bacterial Communities with a Focus on Bacteria Attached to Particles (Doctoral dissertation, The Open University).

Tréguer, P. (2002). Silica and the cycle of carbon in the ocean. Comptes Rendus Geoscience, 334(1), 3-11.

Tréguer, P. J., & De La Rocha, C. L. (2013). The world ocean silica cycle. Annual review of marine science, 5, 477-501.

Tréguer, P., Bowler, C., Moriceau, B., Dutkiewicz, S., Gehlen, M., Aumont, O., ... & Pondaven, P. (2018). Influence of diatom diversity on the ocean biological carbon pump. Nature Geoscience, 11(1), 27-37.

Tremblay, J. É., & Gagnon, J. (2009). The effects of irradiance and nutrient supply on the productivity of Arctic waters: a perspective on climate change. In Influence of climate change on the changing arctic and sub-arctic conditions (pp. 73-93). Springer, Dordrecht.

Tremblay, J. É., Anderson, L. G., Matrai, P., Coupel, P., Bélanger, S., Michel, C., & Reigstad, M. (2015). Global and regional drivers of nutrient supply, primary production and CO2 drawdown in the changing Arctic Ocean. Progress in Oceanography, 139, 171-196.

Tremblay, J. E., Gratton, Y., Fauchot, J., & Price, N. M. (2002). Climatic and oceanic forcing of new, net, and diatom production in the North Water. Deep Sea Research Part II: Topical Studies in Oceanography, 49(22-23), 4927-4946.

Treml, E. A., Halpin, P. N., Urban, D. L., & Pratson, L. F. (2008). Modeling population connectivity by ocean currents, a graph-theoretic approach for marine conservation. Landscape Ecology, 23(1), 19-36. trends in a changing climate. Biogeosciences 11(7):11331–11359.

Tuomisto, H. (2010a). A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. Ecography, 33(1), 2-22.

Tuomisto, H. (2010b). A diversity of beta diversities: straightening up a concept gone awry. Part 2. Quantifying beta diversity and related phenomena. Ecography, 33(1), 23-45.

Turon, X., Antich, A., Palacín, C., Præbel, K., & Wangensteen, O. S. (2020). From metabarcoding to metaphylogeography: separating the wheat from the chaff. Ecological Applications, 30(2), e02036.

Untersteiner, N. (1988). On the ice and heat balance in Fram Strait. Journal of Geophysical Research: Oceans, 93(C1), 527-531.

van de Poll, W. H., Abdullah, E., Visser, R. J., Fischer, P., & Buma, A. G. (2020). Taxon-specific dark survival of diatoms and flagellates affects Arctic phytoplankton composition during the polar night and early spring. Limnology and Oceanography, 65(5), 903-914.

Van Tol, H. M., Amin, S. A., & Armbrust, E. (2017). Ubiquitous marine bacterium inhibits diatom cell division. The ISME journal, 11(1), 31-42.

Vance, D., Little, S. H., de Souza, G. F., Khatiwala, S., Lohan, M. C., & Middag, R. (2017). Silicon and zinc biogeochemical cycles coupled through the Southern Ocean. Nature Geoscience, 10(3), 202-206.

Vandermeer, J. H. (1972). Niche theory. Annual review of Ecology and Systematics, 3(1), 107-132.

Vannier, T., Leconte, J., Seeleuthner, Y., Mondy, S., Pelletier, E., Aury, J. M., ... & Jaillon, O. (2016). Survey of the green picoalga Bathycoccus genomes in the global ocean. Scientific reports, 6(1), 1-11.

Vanormelingen, P., Verleyen, E., & Vyverman, W. (2007). The diversity and distribution of diatoms: from cosmopolitanism to narrow endemism. In Protist Diversity and Geographical Distribution (pp. 159-171). Springer, Dordrecht.

Vellend, M. (2016). The theory of ecological communities (MPB-57). Princeton University Press.

Villa Martín, P., Buček, A., Bourguignon, T., & Pigolotti, S. (2020). Ocean currents promote rare species diversity in protists. Science advances, 6(29), eaaz9037.

Villanova, V., & Spetea, C. (2021). Mixotrophy in diatoms: molecular mechanism and industrial potential. Physiologia Plantarum.

Villanova, V., Fortunato, A. E., Singh, D., Bo, D. D., Conte, M., Obata, T., ... & Finazzi, G. (2017). Investigating mixotrophic metabolism in the model diatom Phaeodactylum tricornutum. Philosophical Transactions of the Royal Society B: Biological Sciences, 372(1728), 20160404.

Villar, E., Farrant, G. K., Follows, M., Garczarek, L., Speich, S., Audic, S., ... & Iudicone, D. (2015). Environmental characteristics of Agulhas rings affect interocean plankton transport. Science, 348(6237).

Villar, E., Vannier, T., Vernette, C., Lescot, M., Cuenca, M., Alexandre, A., ... & Hingamp, P. (2018). The Ocean Gene Atlas: exploring the biogeography of plankton genes online. Nucleic Acids Research, 46(W1), W289-W295.

Violle, C., Navas, M. L., Vile, D., Kazakou, E., Fortunel, C., Hummel, I., & Garnier, E. (2007). Let the concept of trait be functional!. Oikos, 116(5), 882-892.

Volk, T., & Hoffert, M. I. (1985). Ocean carbon pumps: Analysis of relative strengths and efficiencies in ocean-driven atmospheric CO2 changes. The carbon cycle and atmospheric CO2: natural variations Archean to present, 32, 99-110.

Volterra, V. (1928). Variations and fluctuations of the number of individuals in animal species living together. ICES Journal of Marine Science, 3(1), 3-51.

Vorobev, A., Dupouy, M., Carradec, Q., Delmont, T. O., Annamalé, A., Wincker, P., & Pelletier, E. (2020). Transcriptome reconstruction and functional analysis of eukaryotic marine plankton communities via high-throughput metagenomics and metatranscriptomics. Genome research, 30(4), 647-659.

Vyverman, W. (2004). Experimental studies on sexual reproduction in diatoms. Int. Rev. Cytol, 237, 91.

Wassmann, P., Carroll, J., & Bellerby, R. J. (2008). Carbon flux and ecosystem feedback in the northern Barents Sea in an era of climate change. Deep-sea research. Part 2. Topical studies in oceanography, 55(20-21).

Wassmann, P., Kosobokova, K. N., Slagstad, D., Drinkwater, K. F., Hopcroft, R. R., Moore, S. E., ... & Berge, J. (2015). The contiguous domains of Arctic Ocean advection: trails of life and death. Progress in Oceanography, 139, 42-65.

Watson, J. R., Hays, C. G., Raimondi, P. T., MiTarai, S., Dong, C., McWilliams, J. C., ... & Siegel, D. A. (2011). Currents connecting communities: nearshore community similarity and ocean circulation. Ecology, 92(6), 1193-1200.

Webb, T. J., Dulvy, N. K., Jennings, S., & Polunin, N. V. (2011). The birds and the seas: body size reconciles differences in the abundance–occupancy relationship across marine and terrestrial vertebrates. Oikos, 120(4), 537-549.

Weinkauf, M. F., Siccha, M., & Weiner, A. K. (2022). Reproduction dynamics of planktonic microbial eukaryotes in the open ocean. Journal of the Royal Society Interface, 19(187), 20210860.

Whitt, D. B. (2019). On the role of the gulf stream in the changing Atlantic nutrient circulation during the 21st century. Kuroshio Current: Physical, Biogeochemical, and Ecosystem Dynamics, 51-82.

Whittaker, K. A., & Rynearson, T. A. (2017). Evidence for environmental and ecological selection in a microbe with no geographic limits to gene flow. Proceedings of the National Academy of Sciences, 201612346.

Whittaker, R. H. (1960). Vegetation of the Siskiyou mountains, Oregon and California. Ecological monographs, 30(3), 279-338.

Whittaker, R. H. (1965). Dominance and diversity in land plant communities: numerical relations of species express the importance of competition in community function and evolution. Science, 147(3655), 250-260.

Whittaker, R. H. (1972). Evolution and measurement of species diversity. Taxon, 21(2-3), 213-251.

Wickham, H. (2016). ggplot2: elegant graphics for data analysis. springer.

Wieters, E. (2001). Marine macroecology. Trends in Ecology & Evolution, 16(2), 67-69.

Williams, C. B. (1943). Area and number of species. Nature, 152(3853), 264-267.

Wilson, J. B. (2011). The twelve theories of co-existence in plant communities: the doubtful, the important and the unexplored. Journal of Vegetation Science, 22(1), 184-195.

Winter, A., Henderiks, J., Beaufort, L., Rickaby, R. E., & Brown, C. W. (2014). Poleward expansion of the coccolithophore Emiliania huxleyi. Journal of Plankton Research, 36(2), 316-325.

Woodgate, R. A. (2018). Increases in the Pacific inflow to the Arctic from 1990 to 2015, and insights into seasonal trends and driving mechanisms from year-round Bering Strait mooring data. Progress in Oceanography, 160, 124-154.

Woodgate, R. A., Weingartner, T., & Lindsay, R. (2010). The 2007 Bering Strait oceanic heat flux and anomalous Arctic sea-ice retreat. Geophysical Research Letters, 37(1).

Wright, J. L. C., Boyd, R. K., Freitas, A. D., Falk, M., Foxall, R. A., Jamieson, W. D., ... & Dewar, D. (1989). Identification of domoic acid, a neuroexcitatory amino acid, in toxic mussels from eastern Prince Edward Island. Canadian Journal of Chemistry, 67(3), 481-490.

Wright, S. (1949). The genetical structure of populations. Annals of eugenics, 15(1), 323-354.

Xiao, X., Locey, K. J., & White, E. P. (2015). A process-independent explanation for the general form of Taylor's law. The American Naturalist, 186(2), E51-E60.

Xie, W., Yan, Y., Hu, J., Dong, P., Hou, D., Zhang, H., ... & Zhang, D. (2021). Ecological Dynamics and Co-occurrences Among Prokaryotes and Microeukaryotes in a Diatom Bloom Process in Xiangshan Bay, China. Microbial Ecology, 1-13.

Xu, X., Wang, N., Lipson, D., Sinsabaugh, R., Schimel, J., He, L., ... & Tedersoo, L. (2020). Microbial macroecology: In search of mechanisms governing microbial biogeographic patterns. Global Ecology and Biogeography, 29(11), 1870-1886.

Yachi, S., & Loreau, M. (1999). Biodiversity and ecosystem productivity in a fluctuating environment: the insurance hypothesis. Proceedings of the National Academy of Sciences, 96(4), 1463-1468.

Yung, C. M., Ward, C. S., Davis, K. M., Johnson, Z. I., & Hunt, D. E. (2016). Insensitivity of diverse and temporally variable particle-associated microbial communities to bulk seawater environmental parameters. Applied and environmental microbiology, 82(11), 3431-3437.

Zehr, J. P., & Kudela, R. M. (2011). Nitrogen cycle of the open ocean: from genes to ecosystems. Annual review of marine science, 3, 197-225.

Zhang, Y., Xiao, W., & Jiao, N. (2016). Linking biochemical properties of particles to particle-attached and free-living bacterial community structure along the particle density gradient from freshwater to open ocean. Journal of Geophysical Research: Biogeosciences, 121(8), 2261-2274.

Zhou, J., Kang, S., Schadt, C. W., & Garten, C. T. (2008). Spatial scaling of functional gene diversity across various microbial taxa. Proceedings of the National Academy of Sciences, 105(22), 7768-7773.

Zhou, J., Richlen, M. L., Sehein, T. R., Kulis, D. M., Anderson, D. M., & Cai, Z. (2018). Microbial community structure and associations during a marine dinoflagellate bloom. Frontiers in microbiology, 9, 1201.

Zingone, A., Percopo, I., Sims, P. A., & Sarno, D. (2005). Diversity in the genus Skeletonema (Bacillariophyceae). I. A reexamination of the type material of S. Costatum with the description of S. Grevillei sp. nov. 1. Journal of phycology, 41(1), 140-150.