Global Diatom Biodiversity : An Assessment Using Metabarcoding Approach

Shruti Malviya

UNIVERSITE PARIS-SUD

ÉCOLE DOCTORALE : SCIENCES DU VÉGÉTAL

Ecology and Evolutionary Biology Section,
Institut de Biologie de l'ENS(IBENS), CNRS UMR8197, INSERM U1024, Paris, FRANCE

DISCIPLINE : BIOLOGIE

THÈSE DE DOCTORAT

Soutenance prévue le 20/05/2015

par

Shruti MALVIYA

Global Diatom Biodiversity: An Assessment Using Metabarcoding Approach

Composition du jury :

Directeur de thèse :	Chris BOWLER	Directeur de recherche de classe exceptionnelle, IBENS-Paris, FRANCE
Rapporteurs :	Jerome CHAVE	Directeur de recherche, Université Toulouse III- Toulouse, FRANCE
	Guillem CHUST	Senior Researcher, AZTI-Tecnalia, SPAIN
Examinateurs :	Daniel VAULOT	Directeur de recherche de classe exceptionnelle, Station Biologique Roscoff, FRANCE
	Victor SMETACEK	Professeur, Alfred Wegener Institute, GERMANY
President :	Jacqui SHYKOFF	Directeur de recherche, Université Paris-Sud-Orsay, FRANCE

"The footsteps of Nature are to be trac'd , not only in her *ordinary course*, but when she
seems to be put to her shifts, to make many *doublings* and *turnings*, and to use some kind
of art in endeavouring to avoid our discovery."

-Robert Hooke, Micrographia (1665, reprint 2008) , 17.

DEDICATION

To my son, ATHARV,
for being the strongest son and holding “that” special bond.
You have grown into a wonderful 7 year old “good little boy” in spite of your mumma being away from you.

To my husband and best friend, ALOK,
for finding me the light, whenever it was far away.

"Sometimes our light goes out but is blown into flame by another human being.
Each of us owes deepest thanks to those who have rekindled this light."
Albert Schweitzer

ACKNOWLEGEMENTS

I would like to take this opportunity to extend my heartfelt gratitude towards the following people whose contributions have made this endeavor possible.

Dr. Chris Bowler: I feel truly blessed to have worked under your tutelage. You exposed me to different projects, which helped me build a broader perspective for research. I have no words to express my thankfulness for your continual patience, words of encouraging guidance and enthusiastic supervision. The incredibly valuable sessions we spent together shaped my understanding of the subject. Time and again, I felt that either it be my scientific or administrative troubles, once they used to come under your notice, with some magical ease they used to either get sorted or you used to assist me in having the means to do so myself. It has been because of your guidance and motivation that I could steer myself forward through the course of my thesis with confidence. I cannot express in words the amount of reverence I feel for you. I will always be indebted to you.

Dr. Jacqui Shykoff, Dr. Jerome Chave, Dr. Guillem Chust, Prof. Dr. Victor Smetacek, Dr. Daniel Vaulot: for being kind enough to accept the request to be a part of my jury.

Dr. Silvia de Monte: I would like to thank you immensely for all the valuable discussions and suggestions because of these I never had any dearth of ideas to try. I wish I could have learned and imbibed more from you.

Dr. Adriana Zingone: I am extremely grateful for your kind counsel on issues related to diatoms and their ecology. No amount of "thank you" can suffice for your contribution.

Tara Ocean Consortium and all the people I met on board Tara: A massive thank you goes out to *Dr. Colomban de Vargas* and *Stephane Audic* for all their help and support with the data. I earnestly thank all *Tara-nauts* for their kind and extremely helpful discussions. I also owe a thank you to *Dr. Daniele Iudicone* for his valuable advices that has helped in shaping the thesis in a better way. *Eleonora:* I thank you for all your kind help.

MicroB3 Grant: For the project funding that enabled me to have this PhD.

Leila: You have been helping me since before my arrival in France. I can never thank you enough. You are such a warm person and I am immensely grateful to you for all your kindness.

Lucie: You have been my friend first and colleague second. From the very beginning, you have taken those extra measures to make me feel welcome, happy and sane. Thank you for all the

scientific discussions and ideas that you used to lend and for always being so interested in listening to the progress in my work. I will always miss the girl's office fun. I sincerely cherish our friendship and you truly hold a special place in my heart. Thank you for everything.

Flora, Ana, Leandro and Amos: The lows are not too low and the highs are additionally high when one has people like you to share it with. You all have been the most awesome office-mates one could have. I wish we could have spent more time in our office together but whatsoever memories are there, those will always be cherished. Thank you.

Martine: Thank you for all the great times. *Yann*: Thank you for keeping the happy mood high in the lab. *Achal, Alaguraj, Anne-Sophie, Atsuko, Heni, Javier, Omer, Richard, Catherine, Imen, Zhanru* and all other present and past members of the lab: Research is never carried out in isolation, and I would like to thank you for creating an open, friendly environment. It was a privilege that I was a part of such a great team.

Sophie, Camille, Sandrine, and Beatrice: You all have been so kind with your prompt administrative helps. Doing science becomes easier when all such things are well taken care by proficient people like you. Grand merci.

Paris: I came to you knowing not even the language and no person and you gave me some of the best memories ever. Thank you for being so generous in accepting me and thank you for making me meet so many kind hearted people at every step of my stay, some of them ended up being friends for life. Thank you!

Akansha: What should I write for you and how should I thank you? You have been my sister that I never had. You have been my confidant and comforter. I looked forward to returning from lab everyday as I knew you were there. Every day of living in Paris would not have been this good if it was not for you. You made my home away from home and I thank God for letting me find you. Thank you for being there and for keeping me sane and in touch with the 'real world' over the past three years.

And yes now, to *my entire extended family*: I am forever indebted to *my parents* for always being there, to care, protect and raise me, their rightful teachings and unconditional love. *Shrawan*: Though you do not deserve a formal thank you but still thanks for keeping my childhood, alive in me, by continuing to be a prankster and the source of my insane happiness quotient. I wish to thank my entire extended family, especially my *parents-in-law*, for their love and belief in me.

My husband, *Alok* and son, *Atharv*: You have not only seen my dreams through your eyes, but have given me the strength to fulfill those. A list of your contributions would exceed the length of this dissertation and still I will be short of words. Staying away from you people has been the biggest ordeal. But your unconditional support and love made it easier to go through, for me if not always for you both. THANK YOU *Skype* for keeping me somehow and somewhere connected to the epicenter of my world.

And last but not the least, supreme lord, thank you, everything is because of your omnipresent grace!

Table of Contents

CHAPTER 1

Diatom Biodiversity Assessment: General Introduction

Summary

1.1. Diatoms: life in glass houses

"Another way to appreciate diatoms is to realize that they give us every fifth breath, by the oxygen they liberate during photosynthesis." *- David Mann*

General introduction. Diatoms (Class *Bacillariophyceae*) are unicellular, eukaryotic algae, best known for their characteristic silica-based cell wall called a frustule (**Figure 1.1A**). The etymology of the word Diatom comes from the Greek *diatomos*, meaning 'cut in half'. They appeared in the prehistoric era and colonized almost all aquatic environments, from marine to freshwater. They are one of the most important and abundant components of marine phototrophs, contributing roughly 25% of total global primary productivity. The majority is photoautotrophs, but a few have become obligate heterotrophs (Round et al., 1990). They might be solitary or colonial, with some diatoms remaining suspended within the aquatic habitat (i.e. planktonic) while other forms of diatoms are settled within the sediment (i.e. benthic) where they are a major food source for grazing protozoa and animals. Diatoms are abundant and diverse with no accurate count of the number of species, although the highest estimates are around 200,000 extant species spread across all aquatic habitats (Mann and Droop, 1996). They are often referred to as nature's nanofabrication factories because of their outstanding ability to produce complex, beautiful, protective silica frustules that are effectively intricate glass shells (**Figure 1.1B**).

History of diatom research. The history of diatom research goes back more than three hundred years (Round et al. 1990; Flower 2005). The first observation of the diatom dates to 1703, by an unknown Englishman. The work was communicated to the Royal Society of London and published in its Philosophical Transaction (Anonymous, 1703). In the latter half of the 18th century, many diatoms were observed and given classifications. The advent of new technologies and the general availability of electron microscopes, during the late 1980s, further revolutionized the study of diatoms allowing the detailed examination of the ultrastructure of their siliceous frustules (Round et al., 1990). In 1844, Kutzing published the Monograph of 1844 in which he classified all diatoms as algae. Diatoms were also one of the first specimens in which the details of cell division (i.e. mitosis) were examined. The exquisite drawings of diatom mitosis by Lauterborn were published in 1896. During 20th century, many researchers examined diatom species occurrence with respect to environmental factors. From the late 20th century to date, the availability of sophisticated computational tools has permitted the use of diatoms in applied studies.

Applications in ecology. Diatoms are diverse, ubiquitous, and sensitive environmental indicators and, thus, have an enormous ecological importance. With a very large number of ecological-

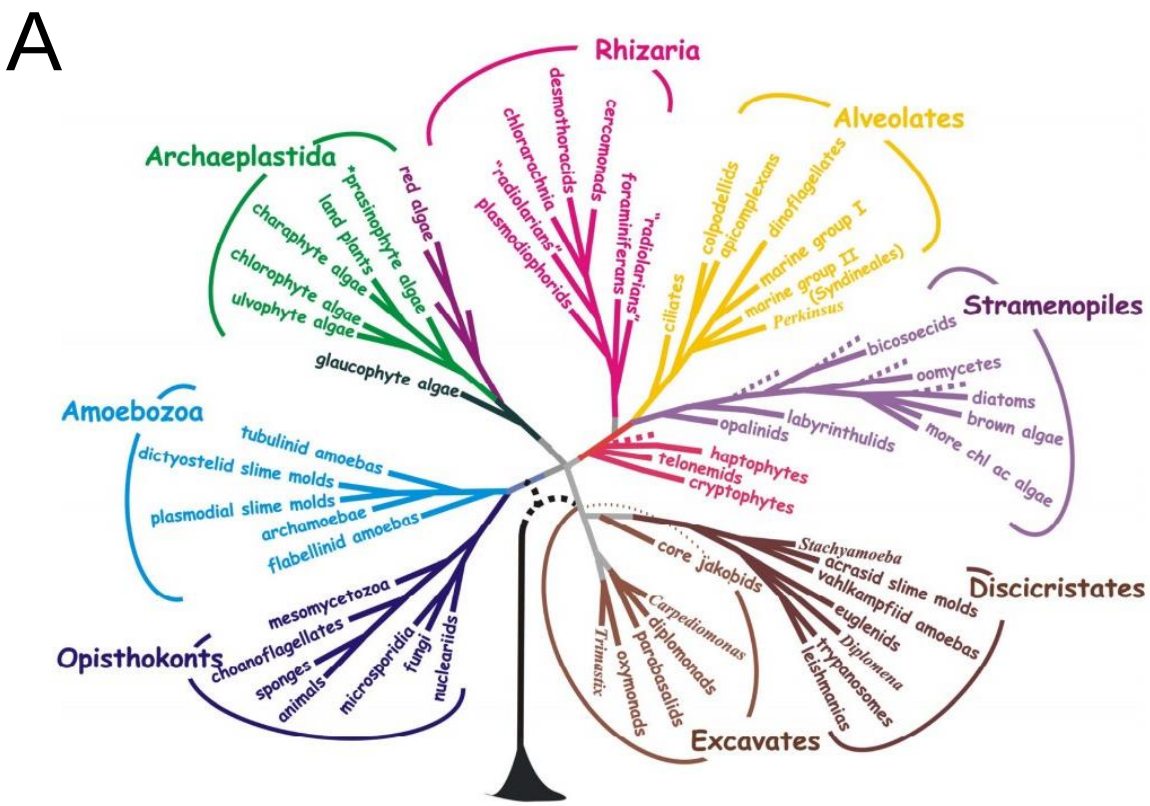


Figure 1.1. (A) Eukaryotic tree of life. The diatoms belong to the stamenopiles supergroup. (Taken from Baldauf, Science, 2003). **(B) *Melosira sp.* (left) and *Lauderia annulata* (right) diatoms collected during the *Tara* oceans expedition (2009-2013).** (courtesy: Christian Sardet, Jennifer Gillette and Chris Bowler).

-ly sensitive species, it is interesting to understand the distributions and occurrences of diatoms for applied ecology. In the recent past, they have been employed in various environmental studies as environmental indicators to study freshwater, marine and brackish ecosystems. They have also been used as indicators in extreme environments, including Arctic and Antarctic waters (Stoermer and Smol, 1999). Moreover, because of their siliceous composition, they are often well preserved in sediments and fossil deposits, making them useful as biogeochemical markers and for archaeological studies.

1.1.1. Diatom ultrastructure

Cytoplasmic features. The diatoms belong to the heterokont algae, i.e. algae with chlorophylls a and c and two flagella of differing sizes (only visible in some species and at specific times during their life histories). The actual protoplast of a diatom is quite similar to that of other algae containing organelles: a nucleus, mitochondria, chloroplasts, Golgi apparatus, endoplasmic reticulum and a large central vacuole. Their nucleus is usually centrally located, migrating to specific sites in the cell as the diatom prepares for cell division (**Figure 1.2**). DNA is often organized as a large number of very small chromosomes. Diatom chloroplasts are characterized by possession of chlorophyll a and c and the primary accessory pigments, β-carotene and fucoxanthin, that give them a characteristic golden color (van den Hoek et al., 1995). The number of chloroplasts and their intracellular arrangement differ among taxa (Cox, 1996), with a typical characteristic feature of heterokonts, having four membranes around them. Cells store energy obtained from photosynthesis in the form of chrysolaminarin and lipids.

Frustule morphology. Diatoms are surrounded by an ornamented compound silica cell wall, called the "frustule", a hallmark of the diatoms (Round et al., 1990). It consists of two overlapping halves known as epitheca (larger upper valve), and a hypotheca (smaller lower valve). The vertical lip or rim of the epitheca is called the epicingulum, and the epicingulum fits over (slightly overlaps) the hypocingulum of the hypotheca. The epicingulum and hypocingulum with one or several connective bands make up the girdle. Many diatoms are heterovalvate, i.e. the two valves of the frustule are dissimilar (**Figure 1.3A**). The frustule elements of diatoms are formed in silica deposition vesicles (SDVs) within the cell under mildly acidic conditions (Vrieling et al., 1999). Calcium binding glycoproteins, called frustulin, coats the frustule and then, they are exocytosed from the protoplast. The form and shape of the frustule is very diverse and species- specific (Round et al., 1990). The evolution of a refined frustule has undoubtedly allowed diatoms to colonize the pelagic oceans and is considered less costly energetically than an organic cell wall (Raven, 1983).

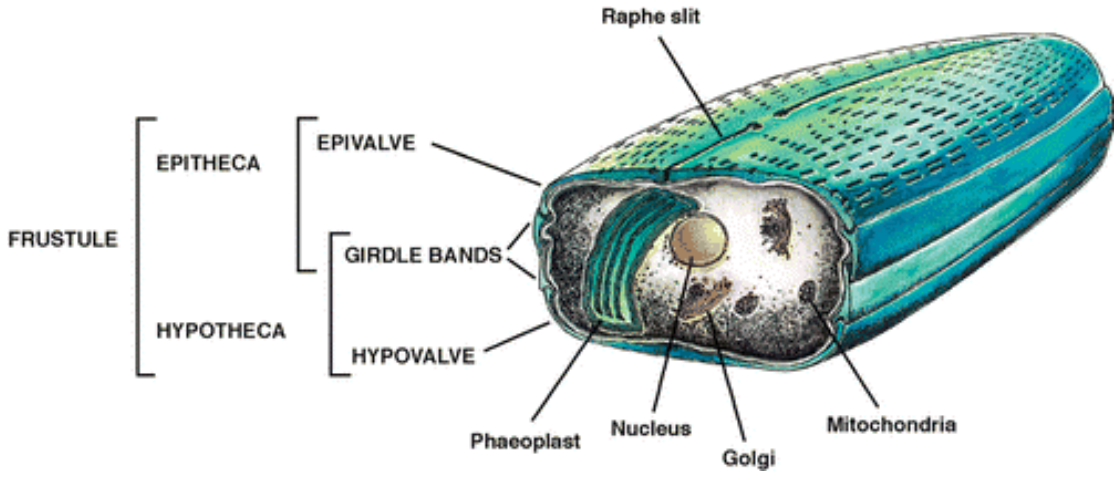


Figure 1.2. Schematic overview of the general structural features of a pennate diatom (Taken from Falciatore and Bowler, 2002).

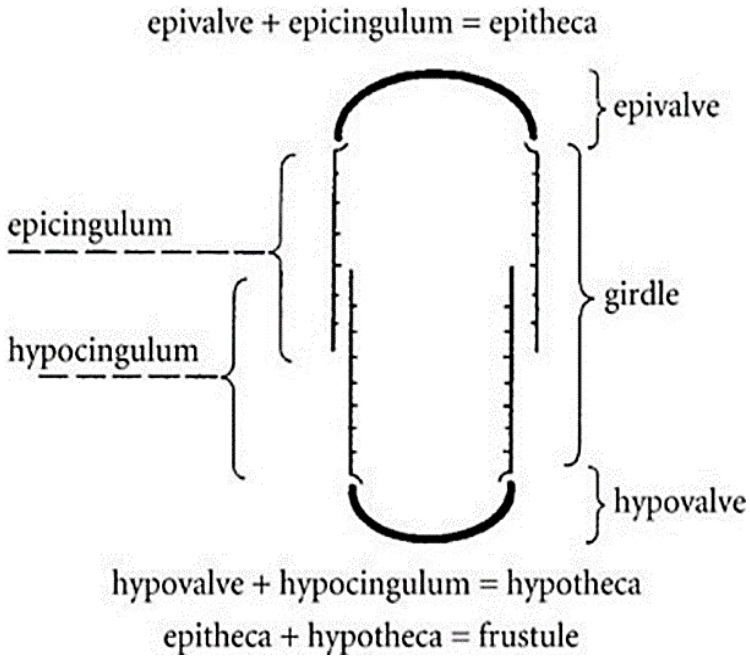


Figure 1.3. Frustule morphology. (Taken from Benten and Harper,2013)

Morphology and classification. Based on the frustule morphology, two main divisions have been recognized, i.e. Centric with round valves and Pennates with more elliptical valves (Round et al., 1990) (**Figure 1.8**). Centric diatoms have valves that are typically circular to elliptical or polygonal in outline. They possess many discoid plastids. Centrics are either classified as radial centrics or polar centrics based on their morphology. Pennates are either araphid pennates or raphid pennates, depending on whether or not they possess a raphe, a longitudinal slit involved with gliding motility. These structural groups have been arranged differently through time. More detailed view on diatom classification is presented later in the **section 1.1.6**.

1.1.2. Habitats and adaptations

As prolific phototrophic organisms, diatoms can live in the open ocean, polar waters, tropical waters, all fresh water areas, soil, snow and even glacial ice. During the course of evolution they have developed different adaptations to survive within each environment. The two main adaptations are (i) storing energy as oils that allow them to be suspended in the water column, and (ii) strong silica frustules that protect them from predation. Additionally, planktonic species often have morphological adaptations that allow them to remain suspended in water. These adaptations to prevent sinking include forming long chains, linked by silica spines (Gersonade and Harwood, 1990). This type of linkage is the most common mode of chain formation and has been seen in radial centric diatoms such as *Paralia*, *Stephanopyxix* and *Aulacosiera*, in the multipolar centrics *Detonula* and *Skeletonema*, in araphid pennates *Staurosira*, *Fragilarioforma* and *Fragilaria*, and even in raphid pennates such as *Diadesmis* (Falkowski and Knoll, 2011). Other diatom species grow attached to surfaces like rocks or aquatic plants, e.g. *Licmorphora* and *Tabularia*. Their frustules are shaped in such a way to aid in attachment. Some species form short stalks, or mucilage pads, while others form long branching stalks, that hold the cells in place and are resistant to waves or high flow in rivers. Apical pads often lead to characteristic zig-zag or stellate (star-shaped) colonies, that resist sinking, found in many araphid pennates and bipolar centrics (Falkowski and Knoll, 2011). Diatoms that have a raphe system (**Figure 1.2**) are able to move over benthic surfaces, whether the surfaces are fine grains of sand, or within the mud of a tidal zone, or even on other diatoms. A few diatoms also form mucilage tubes and move up and down inside the tubes (**Figure 1.4**).

1.1.3. Life history

Diatoms are able to reproduce both sexually and asexually, but primarily by a unique “shrinking division” mode of asexual reproduction (**Figure 1.5A**). During cell division, the two valves get separated, each of them forming the epivalve of the daughter cells and new hypovalves are secreted

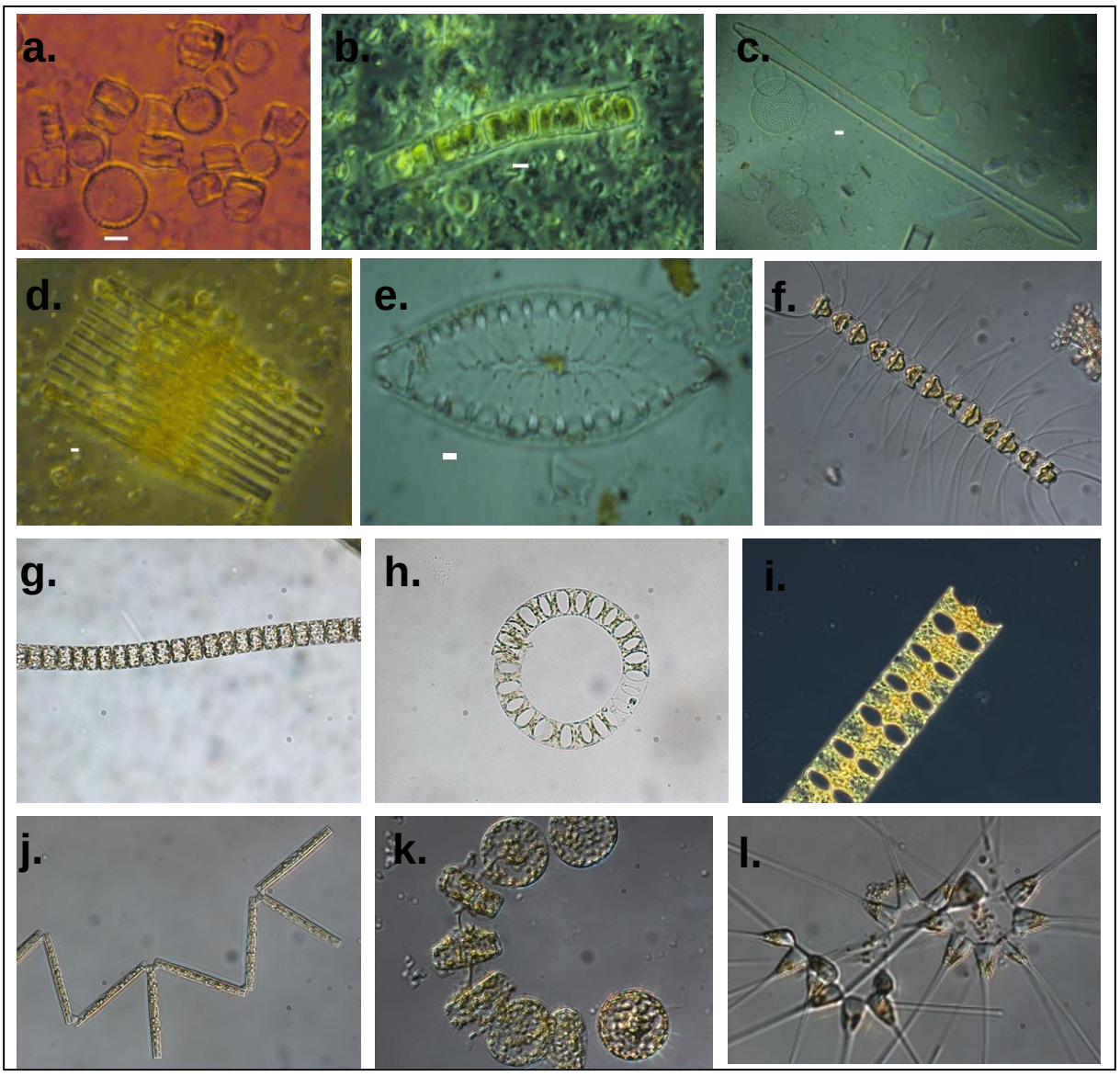


Figure 1.4. Morphological adaptations. (a) *Cyclotella sp.* Usually solitary, but sometimes united in short chains. (b) *Melosira sp.* Cells closely united to more straight, beadlike chains by the middle of the valve faces. (c) *Surirella sp.* Cells solitary. (d) *Tabellaria sp.* Cells quandrangular forming zig-zag or straight filaments. (e) *Synedra sp.* Cells free or united into ribbon-like or star-like colonies. (f) *Chaetoceros sp.* Cells form chains that are coiled, curved or straight. (g) *Detonula sp*. Cells join together in mainly straight, stiff chains by short processes and mucilage threads. (h) *Eucampia sp.* Flattened elliptical cells form spiral, curved chains, joined by flattened apical horns. (i) *Odontella sp*. Heavily silicified cells form curving or spiraling chains, joined by mucous pads on ends of elevations. (j) *Thalassionema sp.* Cells attached together by mucilage pads at their ends into stellate and/or zig-zag-like colonies. (k) *Thalassiosira subtilis* Discoid cells are in chains or are embedded in mucilage. (l) *Asterionellopsis sp.* Cells joined by valve faces into star-shaped or spiraling chains.
[Available from: (a-f) http://msnucleus.org/watersheds/biological/diatomgen.html;
(g-l) http://oceandatacenter.ucsc.edu/]

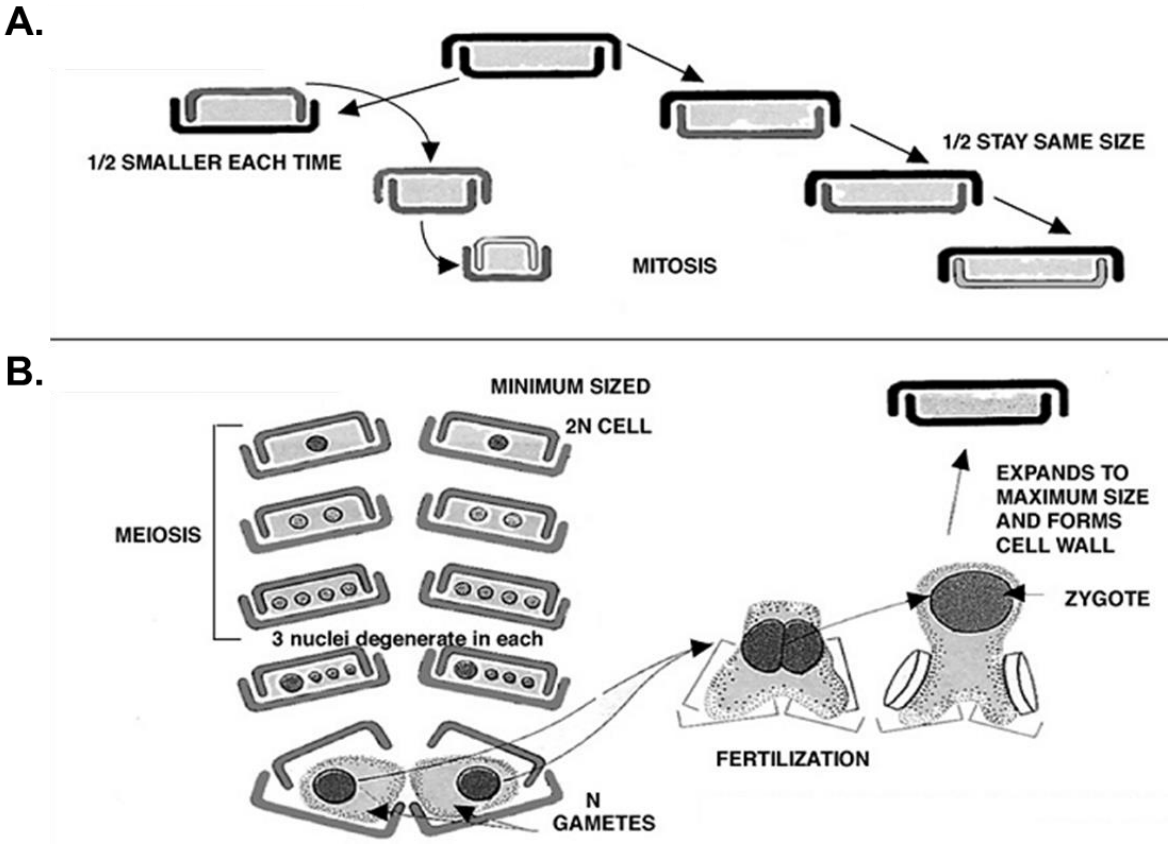


Figure 1.5. Reproduction in diatoms. A. "shrinking division" mode of asexual reproduction. Diatoms cell size progressively decreases with successive generations. **B**. **Sexual reproduction.** When diatom cells shrink to a certain size, they reproduce sexually to form auxospores (credit: Weir et al. 1982).

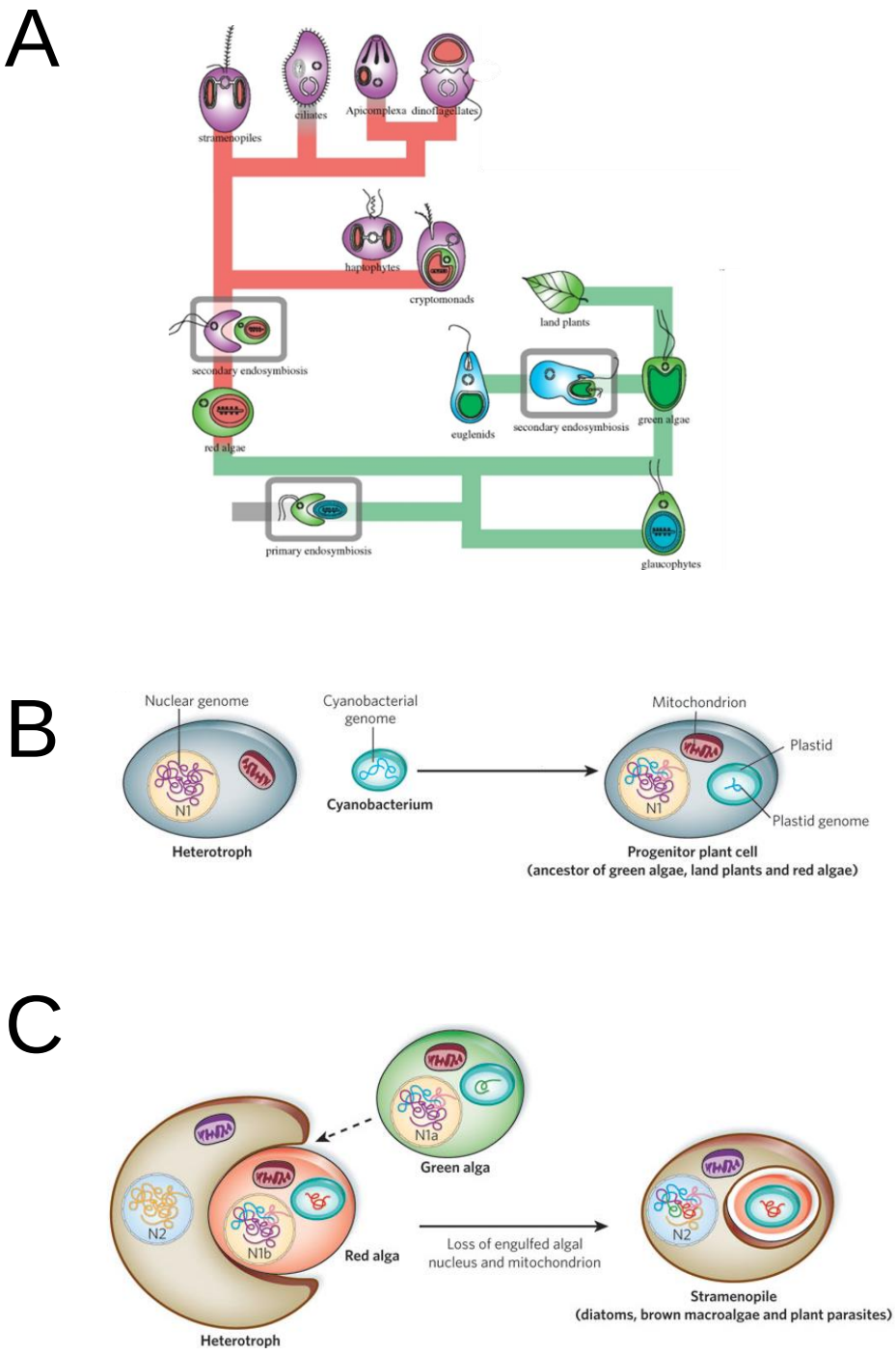


Figure 1.6. (A) Schematic view of plastid evolution. Endosymbiosis events are boxed, and the lines are colored to distinguish lineages with plastids from the green algal lineage (green) or the red algal lineage (red) (adapted from Keeling, 2010). **Detailed representation of the origin of diatom plastids through sequential (B) primary and (C) secondary endosymbiosis.** During primary endosymbiosis, a large proportion of the engulfed cyanobacterial genome is transferred to the host nucleus (N1), with few of the original genes retained within the plastid genome. The progenitor plant cell subsequently diverged into red and green algae and land plants, readily distinguished by their plastid genomes. During secondary endosymbiosis, a different heterotroph engulfed a eukaryotic red alga. Potential engulfment of a green algal cell as well is indicated with a dashed arrow. The algal mitochondrion and nucleus are lost, and crucial algal nuclear and plastid genes (indicated in blue, purple and pink) are transferred to the heterotrophic host nucleus, N2. Additional bacterial genes are gained and lost throughout diatom evolution, but for simplicity this is not indicated here. Based on Armbrust, 2009 and Qui et al., 2013.

within the parent cell (**Figure 1.3**). As the rigid, siliceous cell walls of silica cannot expand, the daughter cells get gradually smaller and smaller. This decrease in cell size with each successive vegetative division continues until it is within a range where environmental parameters may induce sexual reproduction (Edlund and Stoemer, 1997). Gamete formation occurs and they fuse to form a zygote which then gives rise to an auxospore (**Figure 1.5B**). An auxospore possesses a lightly silicified cell wall (perizonia), which allows the cell to expand to its maximum size and then produces a frustule with the normal cell morphology (Kaczmarska et al., 2000, 2011; Wehr and Sheath, 2003). Hence, in diatoms, sexual reproduction is not only a means of inducing genetic variability but also facilitates the enlargement of cells back to their maximum size. In response to stress, i.e. in conditions of low nutrient supply or poor sunlight, diatom cells may form metabolically inactive spores called resting spores (Horner, 2002). Following the onset of favorable conditions, the cells may regain normal functioning.

1.1.4. Secondary endosymbiosis

Diatoms have a complex evolutionary origin and, hence, their genome is called a 'mix-and-match genome' (Armbrust, 2009). Their genomes are the product of a secondary endosymbiosis event in which a heterotrophic eukaryote engulfed a photoautotrophic one and, instead of digesting it, shared with it the ability of photosynthesize (Delwiche, 1999; Bhattacharya and Medlin, 1995). This event led to six different classes of organisms within the Chromalveolate grouping (which incorporates both the Straemnopiles and Alveolates (Fig. 1.1A); Cavalier-Smith, 1999), i.e. Haptophytes, Cryptomonads, Stramenopiles, Ciliates, Apicomplexa, Dinoflagellates (**Figure 1.6A**). This was a sequential event, initially a eukaryotic heterotroph engulfed a cyanobacterium to form the photosynthetic plastids of the Plantae (land plants and red and green algae) (**Figure 1.6B;** Yoon, 2004). This resulted in wholesale gene transfer from the symbiotic cyanobacterial genome to the host nucleus (Reyes-Prieto, 2006).

This primary endosymbiosis was followed by a secondary endosymbiosis event where a different eukaryotic heterotroph captured a red alga (**Figure 1.6C**). Gene transfer continued from the red-algal nuclear and plastid genomes to the host nucleus (Armbrust, 2004). Nosenko and Bhattacharya (2007) identified genes of green algal origin in chromalveolates. This finding led to a hypothesis that they might have originated from an ancient green algal endosymbiont. Later, Moustafa et al. (2009) found the evidence of hundreds of genes of green algal origin in diatoms, supporting the hypothesis that a green alga was involved during the secondary endosymbiosis. Bowler et al. (2008) reported at least 170 red algal genes in the nuclear genome of diatoms, most of which seem to encode plastid components. Owing to this evolutionary history, diatom plastids have been reported to carry out various processes that are characteristic of Plantae plastids, like, photosynthesis, biosynthesis of fatty

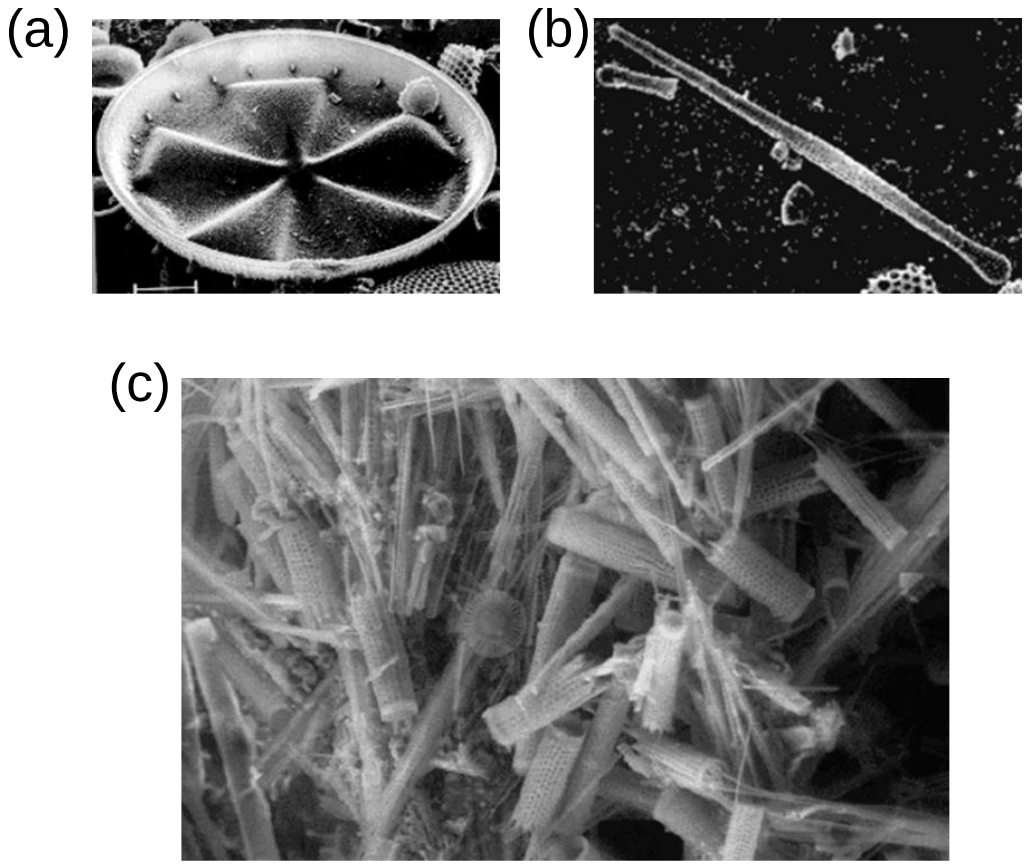


Figure 1.7. Fossil diatoms (a) A single valve from *Actinoptychus heliopelta* and (b) *Sceptroneis caduceus* from marine deposits of Miocene age, the Calvert Formation of Maryland. (c) a lake deposit showing many kinds of diatoms (courtesy: Dr. Karen Wetmore, UCMP Museum).

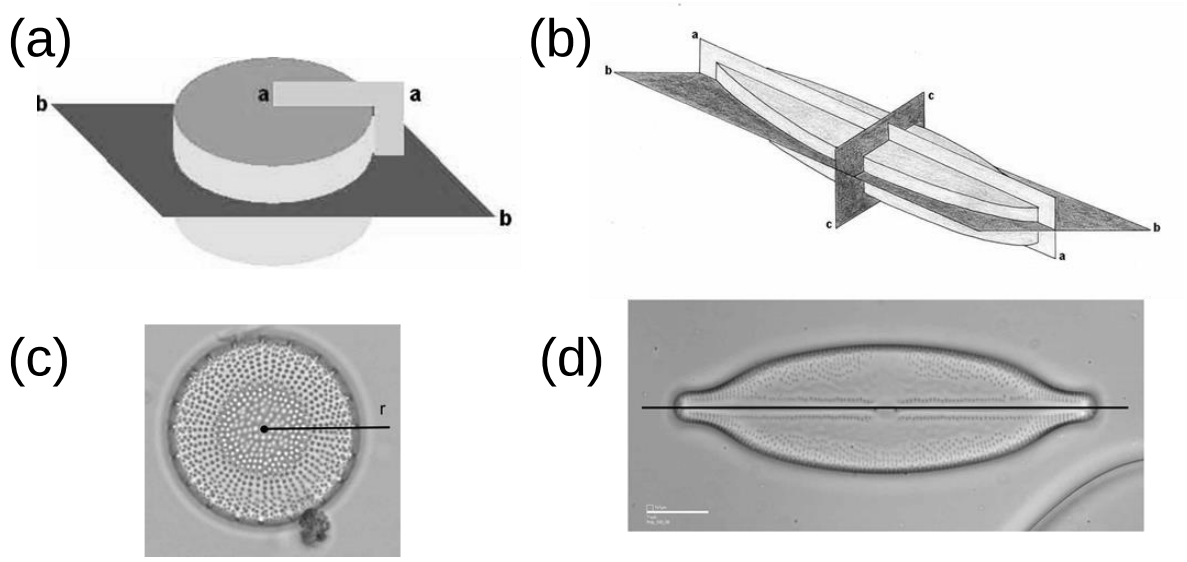


Figure 1.8. Geometrical axis of diatoms. (a) Axes of a centric diatom. a-a radial, b-b valvar, (b) Axes of a pennate diatom. a-a apical, b-b valvar, c-c transapical. (c) A centric diatom showing radial symmetry. (d) A pennate diatom showing bilateral symmetry (courtesy: http://craticula.ncl.ac.uk/).

acids, isoprenoids and amino acids (Armbrust, 2009, Qui et al., 2013). Apart from these, the aforementioned union has sanctified them with a distinctive range of attributes. The most significant is the presence of a urea cycle, which was thought to be restricted to animals (Armbrust, 2004).
It can be speculated that diatoms have a potentially advantageous range of abilities that would not normally be found in a single organism. The silica frustule, thought to be inherited from the exosymbiont, aids in protection from predation, pathogens and desiccation, as well as focusing light into the cell (refs). The subsequent gains (and losses) of specific genes, largely from bacteria, presumably helped them adapt to new ecological niches (Armbrust, 2009). Overall, these processes and many others derived from this unique evolutionary background have ensured their success, making them a highly adaptable group of species with several advantages over other phytoplankton.

1.1.5. Evolutionary and geological history

Fossil records. According to (incomplete) fossil records, the emergence of diatoms took place in the Triassic period (250 Myr ago), although the earliest well-preserved diatom fossils come from the Early Jurassic era (190 Myr ago) (Sorhannus, 2007, Sims, 2006, Armbrust, 2009). The most definitive fossil records for centric diatoms came from the Cretaceous (~145 Myr ago) with the earliest fossil records of araphid (lacking a raphe) pennate diatoms dating from the Late Cretaceous (~145 Myr ago), and raphid pennates from the Middle Eocene (~56.5 Myr ago) (**Figure 1.7**). The earliest freshwater diatoms appeared in the Palaeocene (~65 Myr ago) in Russia and the Late Eocene (~56.5 Myr ago) in North America. However, there are reports of Precambrian (~570 Myr ago) and Triassic (~245 Myr ago) fossils that might be diatoms or diatom relatives (Sims et al., 2006). This belief on diatoms having earlier evolutionary history than expected comes from the property of the silica that it recrystallizes under pressure, which in turn, can destroy diatom fossils. Through the ages, diatom frustules settled down to the bottom of lakes or oceans forming thick deposits of diatomite, or diatomaceous earth. These appear as deposits of white chalky material and are the richest sources of diatom fossils (Benten and Harper, 2013). Diatomaceous earth has a range of commercial applications (see Section 1.1.7).

The rise of diatoms. Following mass extinction in the Cretaceous (~ 65 Myr ago), almost 85% of life was lost, leading to extensive reductions in marine diversity. However, diatoms managed to survive and began to colonize offshore areas, including the open ocean (Armbrust, 2009). Rabosky and Sorhannus (2009) reported that diatom diversity was highest at the Eocene/Oligocene boundary (~30 Myr ago). This era also saw the emergence of raphid pennates, which brought the ability to glide along surfaces and hence expanded the ecological niches greatly (Armbrust, 2009).

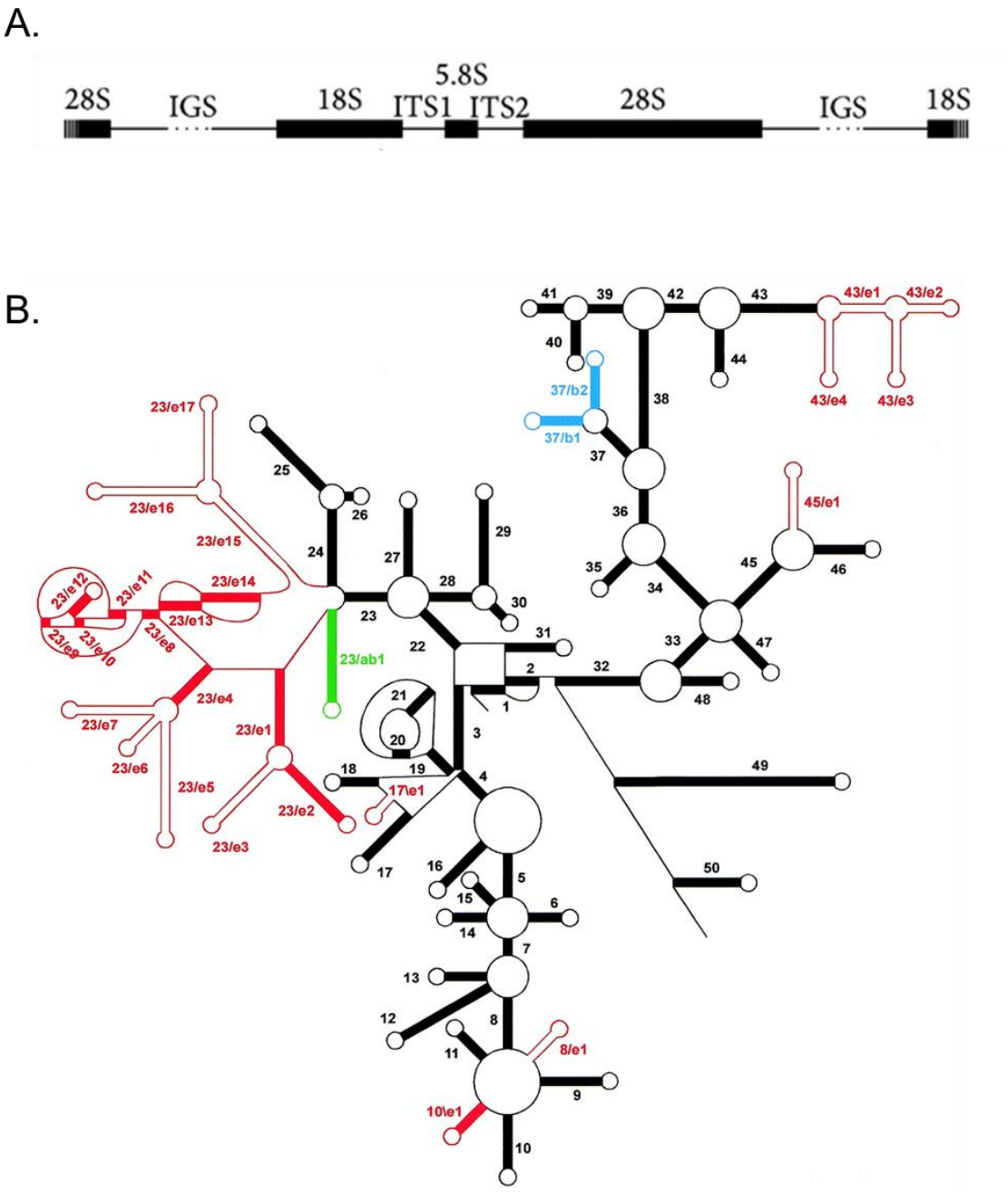


Figure 1.9. Eukaryotic rDNA gene clusters and structure of the 18S rRNA (A) Typical organization of tandemly-repeated rDNA clusters in eukaryotes. 18S, 5.8S, and 28S ribosomal RNA-encoding genes; ITS1 and ITS2 internal transcribed spacers; IGS intergenic spacer **(B) Scheme of the secondary structure of 18S (SSU) rRNA.** The core of the structure common to SSU rRNAs of Archaea, Bacteria and most Eucarya is drawn in black. Helices are numbered in the order of occurrence of their 5′-strand when following the chain from 5′- to 3′-terminus. They bear a different number when separated by a multibranched loop, a pseudoknot loop or a single-stranded area not forming a loop. Bulge loops and internal loops are not shown. Coloured helices are present in Archaea and Bacteria (green), in Bacteria only (blue) or in Eucarya only (red). Those drawn as solid red bars are present in all Eucarya with the exception of the protist taxa Microsporidia, Diplomonadida and Parabasalidea, where some of these helices and even some core helices can be absent. Those drawn as parallel red lines are present only in certain eukaryotic taxa (credit: Wuyts et al., 2002).

1.1.6. Diatom classification

Morphotaxonomy. Diatoms have been classified in different ways by different authors. Round et al. (1990) classified diatoms as a division (Bacillariophyta), whereas Van den Hoek et al. (1995) considered diatoms as a class (Bacillariophyceae or Diatomophyceae) within the division Heterokontophyta. Historically, taxonomists have divided diatoms into two or three major groups, based primarily on the organization of the pattern of striae on the valve. Round et al. (1990) divided the diatoms into three classes: Coscinodiscophyceae, Fragilariophyceae and Bacillariophyceae, which corresponded to three of the main types of valve organization. Informally, these three structural variants can be referred to as 'centrics' (Coscinodiscophyceae), 'araphid pennates' (Fragilariophyceae) and 'raphid pennates' (Bacillariophyceae). Van den Hoek et al. (1995) proposed two major groups of diatoms, Centrales and Pennales. Coscinodiscophyceae and the Centrales are more or less synonymous and are more informally known as the “centric” diatoms. Fragilariophyceae and Bacillariophyceae together correspond to the Pennales and comprise the so-called “pennate” diatoms (**Figure 1.8**).

Molecular phylogeny. In more recent years, molecular markers such as genomic DNA fragments, have been used for phylogenetic analyses to elucidate the evolutionary history of living organisms (Zagoskin et al., 2007). These conserved DNA or RNA nucleotide sequences have enabled researchers, on the one hand, to solve phylogenies at higher taxonomic levels and, on the other, to resolve highly variable sequences to dissect affinities at the species level. One such widely used example of DNA region for reconstructing phylogenies are the genes encoding the ribosomal RNA subunits (rDNAs). The rDNAs encode the RNA components of the ribosome (rRNAs) and form two subunits, the large subunit (LSU) and small subunit (SSU). In most eukaryotes, the 18S rRNA is the small ribosomal subunit, and the large subunit contains three rRNA species (the 5S, 5.8S and 28S rRNAs in mammals, and the 25S rRNA in plants). The rRNA-encoding genes are typically organized in clusters and are separated by internal transcribed spacers (ITS1 and ITS2) and an intergenic spacer (**Figure 1.9A**; Gerbi, 1985).

Owing to the presence of rRNA-encoding rDNAs in all living organisms, rDNA sequences have become a popular choice for molecular taxonomy as it is possible to construct phylogenies for all taxa. The phylogenetic power of rDNA has been repeatedly demonstrated in a wide range of organisms from animals (Freeland and Boag, 1999), including humans (Gonzalez et al., 1990), to higher plants (Alvarez and Wendel, 2003), protists (Sim et al., 2006; Hoshina et al., 2006; Johnson et al., 2007), and fungi (Lutzoni ety al., 2001). Most of the 18S rDNA (region encoding the 18S rRNA) is highly conserved and is generally used for phylogenetic studies at higher taxonomic levels. The tertiary structure of the small subunit ribosomal RNA (SSU rRNA) has been resolved by X-ray crystallography (Yusupov, 2001). The

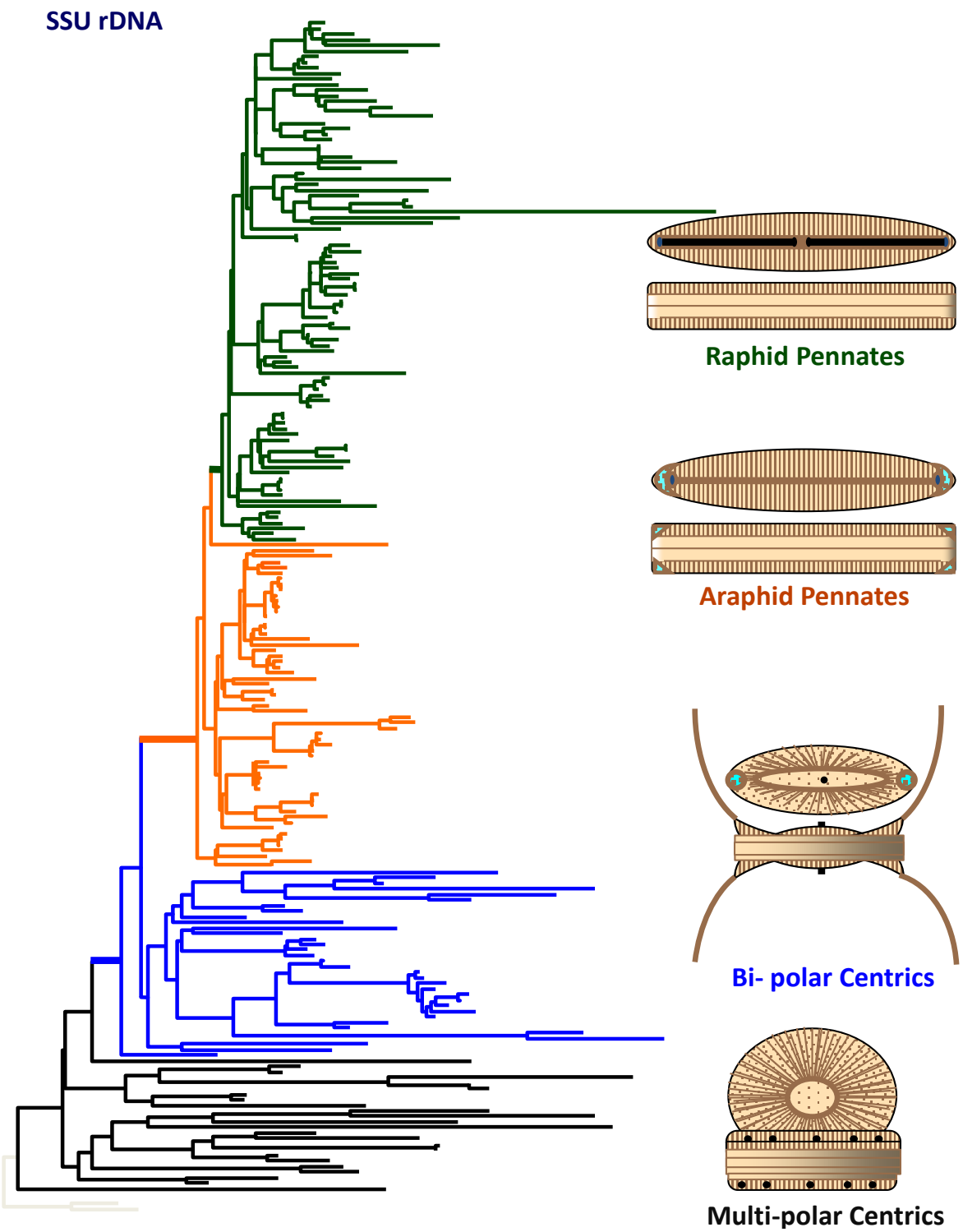


Figure 1.10. Diatom phylogeny inferred from 18S (SSU) rRNA-gene regions of diatoms (credit: Weibe Kooistra, SZN).

secondary structure of SSU rRNA contains 4 distinct domains — the 5', central, 3' major, and 3' minor domains (**Figure 1.9B**).

Kooistra et al. (2003) used 38 diatom SSU sequences and showed "*raphid pennates in a well-supported clade within a paraphyletic araphid group. The pennates as a whole were monophyletic within an apparently paraphyletic group of multipolar centric diatoms. The latter group was essentially composed of a series of clades that collapsed in a polytomy because their basal dichotomies remained unsupported. Pennates and multipolar centrics formed a weakly supported clade, which was sister to radial centrics*".

Based on molecular and morphological data, Medlin & Kaczmarska (2004) proposed a replacement for the traditional view suggesting two new subdivisions (Coscinodiscophytina and Bacillariophytina) for diatoms and a new class, the Mediophyceae, for the bipolar centrics. Adl et al. (2005) adopted Medlin & Kaczmarska's names but treated both the Coscinodiscophyceae and Mediophyceae as paraphyletic taxa (groups that do not include all of the descendents of a single common ancestor).

Theriot et al (2009) proposed a diatom phylogeny based on the nuclear-encoded small subunit of the 18S rDNA gene (SSU) which weakly rejected the classification given by Medlin & Kaczmarska (2004) and others (Sims et al., 2006; Medlin et al. (2008) using parsimony analysis and morphological data. Their results showed that only the Bacillariophyceae (pennate diatoms) were monophyletic, in contrast to Medlin & Kaczmarska (2004) and Sims et al. (2006) who proposed monophyly for each of the Coscinodiscophyceae, Mediophyceae, and Bacillariophyceae.

The diatom phylogeny inferred from 18S rDNA-gene regions of diatoms, shown in **Figure 1.10**, reveals a principal dichotomy leading to a clade of radial centrics (basal clade) and another with multipolar centrics and pennates. The latter exhibited polytomy (a node which has more than two immediate descending branches) containing several clades. The multipolar centrics clade, characterized by a bi-, tri- or multipolar symmetry, cluster in two main polytomic clades. They have been shown to share certain features with araphid pennates, for instance, the ability to produce mucilage from the valve apical pore. They can inhabit both benthic and planktonic environments. Diatoms belonging to the araphid pennate clade constitute a paraphyletic group (i.e. group that does not include all of the descendents of a single common ancestor) consisting of 5 main polytomic clades. Like multipolar centric diatoms, araphid diatoms display a wide range of shapes and life-styles. They are characterized by elongated valves and the ability to form colonies. They can also have both benthic and planktonic

lifestyles. Raphid pennates are the only monophyletic group (i.e. each member is a descendent of a single common ancestor) consisting of cells equipped with the raphe slit. This organelle allows raphid pennates to slide on a substrate and thus, this group abounds in benthic environments. However, many have been reported to have gone back to their planktonic life-styles. In conclusion, it seems that the raphids evolved from araphids and that araphids in their turn are derived from centrics. A similar scenario is supported by the sexual reproduction patterns in these groups, with the evolutionary trend moving from oogamy (ancestral state in centrics) to anisogamy (in pennates) passing through isogamy and imperfect isogamy.

1.1.7. Global importance

Owing to their physiology, diatoms exhibit a remarkable impact on various global phenomena. Their photosynthesis, biogenic silica formation, environmental diversity and a tendency to dominate phytoplankton communities, have led to the major involvement of diatoms in primary production, nutrient cycling, the biological carbon pump, and at the base of the food chain. It has been estimated that diatoms contribute 40–45% oceanic primary productivity, which amounts to 20% of global carbon fixation and oxygen production (Yool and Tyrrell, 2003). Due to the presence of the silica frustule, they play an important role in the biological carbon pump because they facilitate the sinking of organic matter below the photic zone to the ocean floor which provides both essential nutrients for organisms living in the ocean depths and export of carbon to the ocean interior. This makes them a key player in the biological pump as well as the silica cycle.

Diatom growth is limited by factors such as nutrient availability; however, when there are large nutrient influxes or seasonal changes, diatoms can form large blooms up to several square kilometers in size. When the nutrients, principally nitrate and silicate, get depleted, the bloom dissipates, forming aggregates of silicified cells that sink to the ocean floor. This accumulation of diatom frustules in sediment forms diatomaceous earth, a material that is used in a number of industrial applications such as abrasive pastes, water filters, fillers, insulators, and (non-toxic) insecticides (Mann, 1917). Furthermore, most petroleum deposits are derived from diatoms that have sedimented to the seafloor over geochemical time scales (Denman, 2008). The structural and physical properties of the frustule are the focus of several research areas into nanotechnological applications. These include drug delivery solar technology, microfluidics, catalyst production and bio-sensing. Lipid production in diatoms is also drawing interest as a source of renewable oil. Although the global contributions made by diatoms are already significant, these technologies may be the key to drawing the public eye to the importance of these microscopic algae and the roles they play for the well-being of our planet.

In general, diatom species are very particular about the water chemistry in which they live. Some species have distinct ranges of pH and salinity where they will grow. Diatoms also have ranges and tolerances for other environmental variables and different types of human disturbance (Birks, 2010; Smol and Stoermer, 2010). As a result, diatoms are used extensively in environmental assessment and monitoring. Other applied uses of diatoms include oil and gas exploration, forensics, as indicators of atmospheric transport and more broadly, applied research concerning geological, biological, and climatic processes.

Furthermore, because the silica cell walls do not decompose easily, diatoms in marine and lake sediments can be used to interpret conditions of the past. Paleoecology is a field that utilizes both living and sub-fossil diatom valves that are preserved in marine and freshwater sediments to understand the environmental factors influencing the modern presence and abundance of the species. Scientists thereby apply this knowledge of species preference in modern conditions to interpret the diatom species from the past, and the historical conditions that prevailed when they were alive. Thus, diatoms constitute successful single-celled organisms that diversified to fill multiple niches, outcompeting other phytoplankton to take a key position in driving global biological and biogeochemical processes.

1.2. Marine biodiversity and biogeography

A general definition of biodiversity is "the collection of genomes, species, and ecosystems occurring in a geographically defined region" (NRC, 1995).

1.2.1. General introduction

Biodiversity is a contraction of 'biological diversity' and can be defined as the "*full variety of life on Earth*" (Takacs, 1996). It is the study of the processes that create and maintain variation and takes into account the variety of individuals within populations, the diversity of species within communities, and the range of ecological roles within ecosystems. It has three components: species diversity, ecosystem (or habitat) diversity, and genetic diversity. High diversity has often been perceived as a synonym to ecosystem health, as diverse communities are believed to have increased stability, increased productivity, and resistance to invasion and other disturbances.

Biodiversity and ecosystem functioning. Processes central for the functioning of ecosystems might be maintained by several or very few organisms, which leads to the question whether there exists any relationship between biodiversity and ecosystem functioning. The answer to this becomes relevant when environmental conditions result in the loss of biodiversity, as is currently believed to be the case at a global level (Loreau et al., 2002). For the past two decades, ecologists have been arduously trying to describe and quantify the effects that biodiversity can exert on the various processes within ecosystems. Through these studies, it has been found that changing diversity has profound effects on primary production, nutrient retention, and ecosystem stability (Chapin et al., 2000). As our understanding of the relationship between biodiversity and ecosystem functioning develops, conservation and management efforts should benefit.

The recent advances made in functional biodiversity research have led to a new synthetic ecological framework, which has even been denoted as a new paradigm of ecology. While biodiversity has historically been seen as a response variable that is affected by climate, nutrient availability and disturbance, this new emerging paradigm, called '*Biodiversity-Ecosystem Function Paradigm*' (Naeem, 2002), explains the environment primarily as a function of diversity, underlining the active role of the biota in governing environmental conditions. It does not deny, of course, the influence of the environment on organisms. More specifically, within this framework, a specific ecosystem function is thus seen as a combined influence of (i) biodiversity and the functional traits of the organisms involved, (ii) associated biogeochemical processes, and (iii) the abiotic environment. This '*Biodiversity Ecosystem-Function Paradigm*' has shifted the scientific perception of diversity towards the

biodiversity being the driver of ecosystem functions. Based on a broad array of studies from marine, terrestrial and freshwater ecosystems, Hillebrand and Matthiessen (2009) summarized that "(1) losing diversity in an assemblage tends to reduce ecosystem process rates mediated by this assemblage, e.g. the production of organic biomass and the efficiency of resource use (Hooper et al. 2005; Balvanera et al. 2006; Cardinale et al. 2006), (2) both effects become stronger over time (Cardinale et al. 2007; Stachowicz et al. 2008), and (3) loosing diversity also affects certain (but not all) aspects of stability (Hooper et al. 2005; Balvanera et al. 2006)".

Marine biodiversity. Extending over three quarters of the surface of the Earth, oceans are a precious asset. Life likely originated there and they support a large share of global biological diversity. Marine ecosystems play a key role in global biogeochemical cycles and patterns of weather and climate. Economically, there is considerable reliance on the world's oceans – from fisheries which support over 15% of the global protein supply, to off-shore petroleum production, along with the millions of jobs supported by tourism and fishing. Indispensable to life itself, the marine environment is facing direct and indirect impacts from human activities. A wide range of threats such as increasing acidification, coral bleaching, toxins and chemical pollution, nutrient overloading, and fisheries depletion including many others, are undermining the ocean's ability to sustain ecological functions. Marine debris, made up of persistent, manufactured solid materials such as plastics, is another major growing concern as they are discarded in the marine and coastal environment and persist almost indefinitely. All these factors together have resulted in the loss or degradation of marine biodiversity.

1.2.2. Characterizing biodiversity

Whittaker (1972) described three terms for measuring biodiversity over spatial scales (**Figure 1.11**), namely,

(a) *Alpha (α) Diversity.* It refers to the diversity of the community within one site (or one sample), i.e., the number of species and their proportion within one sampling site (Whittaker 1960, 1967). Some commonly used indices to describe alpha diversity include Richness, Shannon's index (H), Simpson's index (D) and Renyi entropy.

(b) *Beta (β) Diversity.* It is defined as the difference in species composition between communities. Higher beta diversity implies low similarity between species composition of two communities and is usually expressed in terms of similarity index between communities in the same geographical area (Whittaker 1960, 1967). Some commonly used beta diversity indices include Bray-Curtis dissimilarity, percent similarity index (PSI), and Jaccard's index (qualitative index).

(c) *Gamma (γ) Diversity.* It is a measure of the overall diversity for the different ecosystems within a region (Whittaker 1960, 1967). Hunter (2002) defines gamma diversity as "geographic-scale

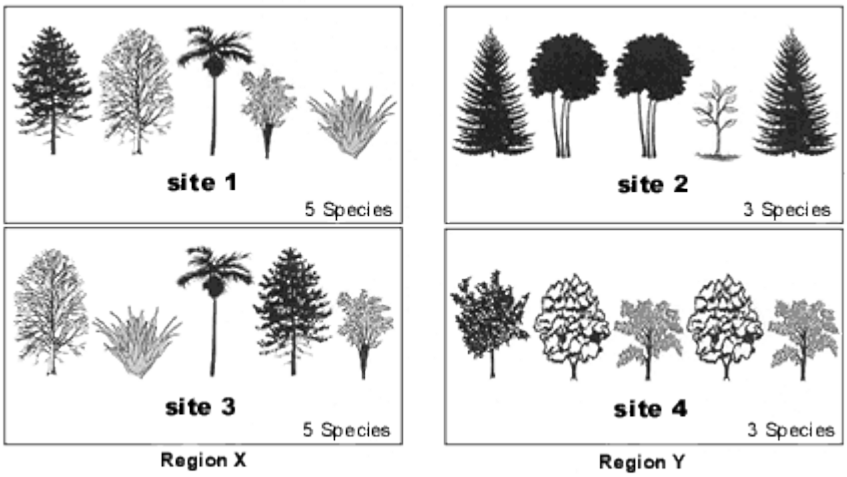


Figure 1.11. Diagram showing biodiversity expressed at several scales. Alpha diversity is measured locally, at a single site, as at sites 1 and 2. Site 1 has higher alpha-diversity than site2. Beta diversity measures the amount of change between two sites or along a gradient, as in regions X and Y. Region Y has higher beta-diversity than region X, as there is a higher turnover of species among sites in region Y. Gamma diversity is similar to alpha diversity, only measures over a large scale. Both alpha- and beta-diversity contribute to alpha diversity. Region X has high alpha-diversity at its sites, but they are all fairly similar; the region thus has low beta- diversity and only moderate gamma diversity. Region Y has low alpha-diversity at its sites, but the sites differ from each other; the region therefore has high beta-diversity, and higher gamma-diversity than region X. (Taken from Figure 5.6 in Perlman and Adelson, 1997).

species diversity".

Several indexes and quantitative measures of biodiversity have been developed. The simplest approach is to express diversity as the number of species on a site or in a community (species richness). Shannon-Wiener's Index (H') and Simpson's Index (λ) are the most commonly used indexes of diversity in ecological studies (detailed description in Appendix B). However, ecologists have suggested that biodiversity measures should be interpreted carefully. As indicated by various studies, some habitats are stressful and so few organisms are adapted for life there, but, those that do, may well be unique or, indeed, rare. Such habitats are important even if there is little biodiversity.

1.2.3. Microbial biogeography: processes and patterns in microbial diversity

> *"… there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns - the ones we don't know we don't know"* *-D. Rumsfeld*

Biogeography is a science that attempts to describe and explain spatial patterns of biological diversity and how these patterns change over time (Ganderton and Coker, 2005; Lomolino et al., 2006). In other words, biogeographers seek to answer the seemingly simple question: Why do organisms live where they do? There has been a recent interest in microbial biogeography. However, the existence of microbial biogeography has been questioned ["*There is no biogeography for anything smaller than 1 millimeter*", Bland Finlay quoted in Whitfield (2005)"]. In contrast to this report, several studies have reported that many microbial taxa exhibit biogeographical patterns; microbial communities are not homogeneous across habitat-types, and within a given habitat, microbial diversity can vary between locations separated by millimeters to thousands of kilometers. However, ecologists agree that it is often complicated to understand and document microbial biogeography. Longhurst et al. (1995) partitioned the world ocean into provinces based on the prevailing role of physical oceanographic contextual data as a regulator of phytoplankton distributions. The four principal biomes defined were Westerlies, Trades, Polar and Coastal Biomes (**Figure 1.12**). These were further subdivided into 52 provinces based on measured and satellite-derived data (Longhurst et al., 1995; 1998; 2006). This partition has been used to systematically classify discrete oceanographic provinces providing a mechanism for enhanced comparative analysis of ecosystem processes, community composition, organismal biogeography and trait attributes (Oliver and Irwin, 2008; Gomez-Pereira et al. 2010; Brown et al., 2014).

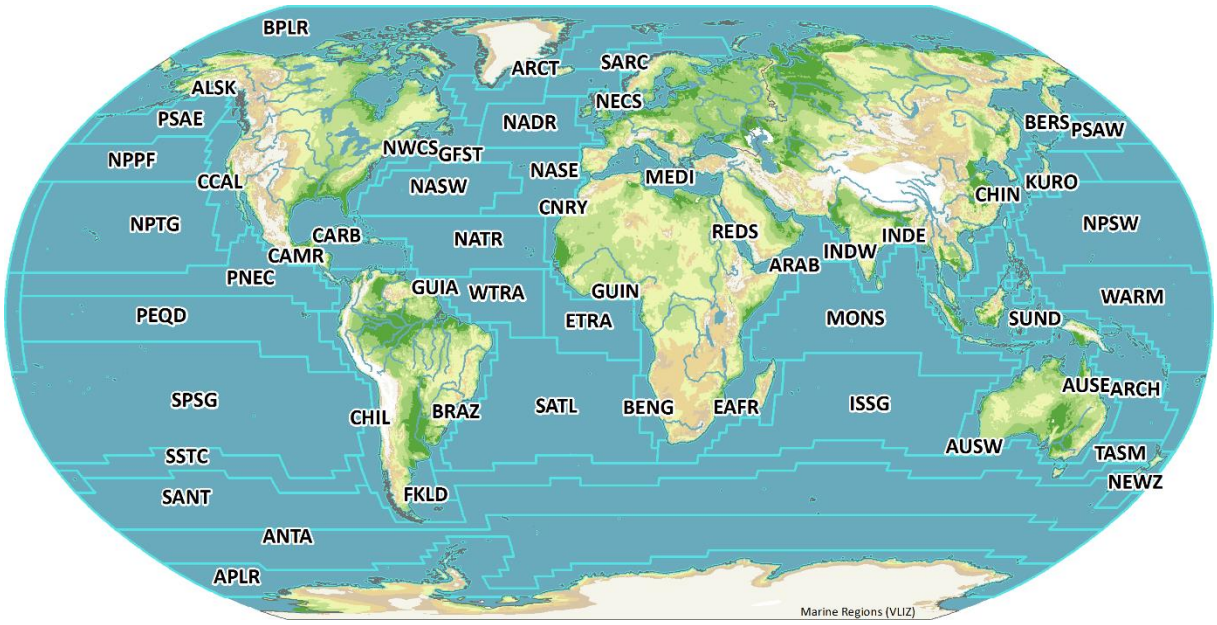


Figure 1.12. Longhurst Biogeographical Provinces. The figure shows the partition of the world oceans into provinces as defined by Longhurst (1995; 1998; 2006), based on the prevailing role of physical forcing as a regulator of phytoplankton distribution. Four principal biomes recognized were: the Polar Biome, the Westerlies Biome, the Trade-Winds Biome, and the Coastal Boundary Zone Biome. These four Biomes are recognizable in every major ocean basin. At the next level of reduction, the ocean basins are partitioned into provinces, roughly ten for each basin. (Available from: http://www.marineregions.org/).

1.2.3.1. Processes regulating microbial biogeography

Historical processes. Species richness in a given location is the result of three factors, i.e. the rate of speciation, the rate of extinction, and the dispersal of species from other locations. Dispersal has been reported to be one of the key processes shaping microbial biogeography, however, its extent has been a topic of debate for long. Finlay (2002) has argued that any organism less than 1 mm in size is likely to be ubiquitous due to an essentially unlimited capacity for long distance dispersal simply by chance (Finlay, 2002; Fenchel, 2003; Martiny et al., 2006). But at larger spatial scales, the active dispersal of microbes can vary considerably between microbial taxa owing to mode of transport, habitat characteristics, population densities, and the ability of the microbe to survive the transport process itself (Jenkins et al., 2007; Martiny et al., 2006). Thus, a dispersal combined with the ability to survive long distance transport may result in few geographic constraints on microbial biogeography.

Processes in variable environments. A comprehensive understanding of the factors that generate biogeographic patterns requires a complete understanding of how the current and past environment shapes dispersal, speciation and extinction. The idea that environmental heterogeneity drives biogeographic patterns is best summarized by the Baas Becking hypothesis "*everything is everywhere, but, the environment selects*" (Baas Becking, 1934; de Wit and Bouvier, 2006). In other words, this hypothesis proposed that there is effectively no limitation to dispersal, and that biogeographic patterns solely reflect contemporary environmental conditions. Thus, similar environments will harbor similar microbial taxa regardless of the geographic distance between the environments.

From an oversimplified perspective, there are two general factors that may contribute to the formation of biogeographical patterns: dispersal limitation and environmental heterogeneity (Lin et al., 2013). In the past few years, there has been an inclination towards the consideration that dispersal alone is unable to alter biogeographical patterns unless it is accompanied by successful establishment (or colonization), which is under the influence of a wide variety of biotic and abiotic processes. Recent studies have indicated that both environmental heterogeneity and dispersal limitation (i.e. history) have relative roles in driving spatial variation in microbial communities.

1.2.3.2. Characteristic patterns of biodiversity

The patterns of diversity, in general, are governed by the combinations of three factors, i.e. random processes of birth, death and migration, history, and necessity. Recent methodological advances have made it possible to survey a large portion of the microbial diversity on Earth and to quantify the biogeographical patterns exhibited by microbes living in a wide range of environments. As these techniques and methodologies continue to improve at a nearly exponential rate, the field of microbial

biogeography is poised for significant advances. Considering the fact that "microbes" represent a broad array of taxa (i.e., bacteria, fungi, archaea, viruses, and protists), it is unlikely that all microbial taxa will follow a common set of concepts and patterns. Gaston (2000) reviewed a series of broad-scale (geographical) spatial patterns well-documented in the literature, to describe the heterogeneous distribution of biodiversity. These included (i) latitudinal gradients in species richness (Clarke and Crame, 1997; Stevens, 1989; Gaston, 1996; Roy et al., 1998), (ii) species-energy relationships (Roy et al., 1998; Turner et al., 1987; Rutherford et al., 1999), and (iii) relationships between local and regional richness (Griffiths, 1997; Cornell, 1999; Ricklefs and Schluter, 1993). It has been reported that the patterns observed may vary with spatial scale, e.g. that processes operating at regional scales can influence patterns observed at local ones and, most importantly, that patterns in biodiversity are unlikely to have a single primary cause (Gaston, 2000).

1.2.4. Future directions in the study of microbial biogeography

The statement by E.O. Wilson that "*microbial diversity is beyond practical calculation*" (Wilson, 1999) is likely to be accurate in many environments and for a variety of microbial taxa. Nonetheless, the past decade has seen rapid advances in the field of microbial biogeography and biodiversity with the emergence of new tools and methods that have provided us with unprecedented abilities to study microbial communities. There is also a growing recognition that microbes do exhibit biogeographical patterns and that, by studying these patterns, we may be able to develop biogeographical theories and hypotheses that apply across the entire tree of life (Fenchel et al., 1997; Finlay 2002; Hedlund and Staley, 2003). However, it is important to recognize that the "*unknown unknowns*" and "*known unknowns*" in microbial biogeography currently outnumber the "*known knowns*". The study of microbial biogeography will assist us in building a predictive understanding of microbial diversity and the factors influencing this diversity across space and time. Microbiologists may be able to use pre-existing concepts in biogeography to understand microbial systems, or we may find that such concepts, which are largely derived from studies of plants and animals in terrestrial environments, are not directly applicable. Either way, the incorporation of microbiology into the field of biogeography promises to be a fruitful endeavor as many fundamental questions remain unanswered. By studying microbial biogeography, we will move closer to understanding the full breadth of biological diversity on Earth.

1.3. Metabarcoding: a new paradigm for biodiversity assessment

Basic approach. Paul Hebert recognized that a small strand of DNA contains enough information to identify millions of species (Hebert et al., 2003). His technique for species identification proved to be successful and numerous studies have shown that DNA barcoding can be used effectively in several groups (Hajibabaei et al., 2008). The basic idea behind DNA barcoding is that new species can be identified by comparing (part of) their DNA to DNA from other species (**Figure 1.13**). This reference DNA could be collected in for instance a reference library. If the DNA (i.e. the barcode) of the target species differs enough from the reference DNA, the target species could be considered as a new or different species. When a target species is collected, its barcode region first needs to be amplified and sequenced. The barcode is then compared to the reference library to determine if the barcode belongs to a "new" species (which was not yet present in the reference library) or if the barcode belongs to a species already present in the library.

DNA barcoding is an approach where a short fragment of a conserved DNA fragment (400 - 1000 bp) from a standardized region of genome is used for species identification and discovery, very similar to the barcodes used in supermarkets to equate products with prices. It has been used as a method for establishing a correlation between taxonomically undetermined individuals to a taxon with similar genetic sequence in a given reference database (Ratnasingham and Hebert, 2007). Initially, DNA barcoding was mainly focused on taxonomic research. However, recent advances in next generation sequencing (NGS) have advanced the dimensions of DNA barcoding, which has been pivotal for both basic and applied research (Nagy et al., 2013).

Ideal barcode. Valentini et al. (2009) summarized five criteria for an ideal barcoding system, which are as follows:

(a) The gene region sequenced should be nearly identical among individuals of the same species, but different between species,
(b) It should be standardizable, with the same DNA region used for different taxonomic groups,
(c) The target DNA region should contain enough phylogenetic information to easily assign,
(d) Unknown or not yet 'barcoded' species to their taxonomic group (genus, family, etc), It should be extremely robust, with highly conserved priming sites and highly reliable DNA amplifications and sequencing,
(e) The target DNA region should be able to allow amplification from environmental DNA.

Potential strengths of DNA metabarcoding in ecological studies. Ecological studies often need syste-

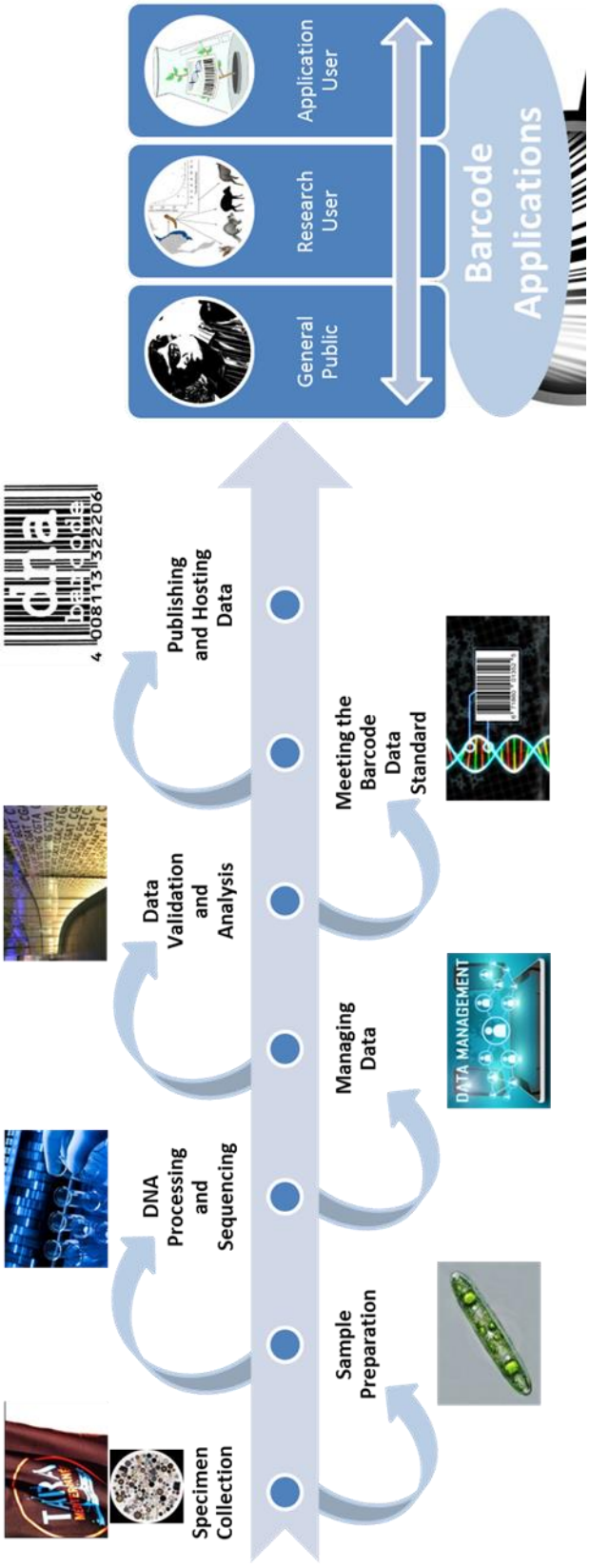


Figure 1.13. Metabarcoding: basic approach

-matic determination of the species carrying out ecological processes. Obtaining such data using traditional methods requires a significant expertise and is often time consuming. The recent development of DNA-based barcoding for species identification has drastically simplified this identification step (Hebert et al. 2003a). DNA metabarcoding couples the principles of DNA barcoding with NGS technology. As much as Earth's unexplored wilderness shrinks every day, we are actually discovering more species than ever before. By doing more systematic and thorough surveys, by heading deeper into unknown territory, and by using advanced tools like DNA barcoding, we are uncovering new species at a record rate.

Marine microbial eukaryotes are among the most diverse groups on the planet with estimates ranging from 500,000 to nearly 10 million species (Appeltans et al., 2012). But these are vastly under documented. The present rate of reliable identification and description is certainly inadequate for the task of describing all species, and new discoveries are often the result of unexpected encounters. Diatoms are present in all types of water bodies and their species diversity is influenced by environmental conditions. Morphological identification of diatoms beyond genus level is difficult and, therefore, DNA barcoding stands out to be one of the potential techniques which can help with accurate identification of species up to taxa level. However, species characterization needs sequences with high discriminatory power. Various gene regions proposed as barcode markers for diatoms include : (1) Mitochondrial cytochrome oxidase I gene (cox1) (Evans 2007, 2008; Saunders 2005; Blaxter 2004; Blaxter et al., 2004); (2) Chloroplast ribulose-1,5 bis-phosphate carboxylase oxygenasegene (rbcL) (Saunders 2005, 2008); (3) Combination of nuclear 5.8S rRNA gene and ITS2 (Moniz and Kaczmarska 2009, 2010); (4) Nuclear small ribsomal subunit (SSU-rRNA gene) (Behnke et al., 2004, Sorhannus 2007; Jahn et al., 2007); (5) V4 region 18S locus - this is 390 - 410 bp long fragment of 1800 bp long 18S rRNA gene locus which represents the highly variable and most complex region within 18S locus (Zimmerman et al., 2011).

Other applications. The possibility of identifying all the species present in an environmental sample opens doors to new approaches in ecological research. It can be a good addition to the already used taxonomic methods instead of a replacement. Indeed, an integrative approach is being more and more widely adopted by many scientists in recent years (e.g., Hajibabaei et al. 2008). In the early days, DNA barcoding focused mainly on taxonomic research. Development of NGS technology has extended the application of metabarcoding in various ways, including:

(a) Conservation biology for biodiversity surveys and to find out the traces of nearly extinct species (Stoeckle 2003; Taberlet 2012a; Callaway 2012; Valentini et al. 2009),

(b) Evaluation of interactions between species and diet analysis (Stoeckle 2003; Taberlet 2012a;

Valentini et al. 2009),

(c) Investigation of changes in species distribution and niche stability (Taberlet et al. 2012a),

(d) Applications in the area of 'biosecurity' due to the potential to accurately identify invasive species (Armstrong and Ball, 2005; Valentini et al. 2009),

(e) Biomonitoring - to monitor illegal trade and by-products (Valentini et al. 2009),

(f) Toxicology (harmful species)

(g) Reconstruction of phylogenetic structure

Limitations. Metabarcoding is a relatively new research field and a lot of research is being conducted to improve the methods. The metabarcoding approach seems very promising for the future, but in order to achieve those promises, some gaps need to be filled and limitations have to be overcome. Some of the limitations associated with metabarcoding include:

(a) Sampling strategy

(b) Universal barcodes

(c) Reference libraries

(d) Quantification of results

(e) Other limitations related to PCR and DNA sequencing

(f) Bioinformatics tools for analyzing data

In spite of associated limitations, metabarcoding appears to be one of the most promising research areas, like biodiversity monitoring, animal diet assessment, reconstruction of paleo communities, and worthy of extensive exploitation to evaluate its full potential (Taberlet et al. 2012c). The advancement of NGS technology and its cost reduction has made metabarcoding even more practical.

Estimates of total numbers of species of eukaryotes. A recent estimate of total number of eukaryotic species suggests that there are ~8.7 million species on our planet. With only 1.2 million species catalogued (Mora et al., 2011), more than 80% of species remain undiscovered. So far, only a small fraction of species on Earth (~14%) and in the ocean (~9%) have been indexed (Mora et al., 2011). With an increasing extinction rate and the current description rates of eukaryote species, it can be speculated that many of the species will become extinct before we know they even existed. Also, previous studies have reported that the species which have been catalogued are those with higher geographical range and abundances. This suggest that majority of undiscovered species are small ranged and concentrated in hotspots and less explored areas such as the deep sea and soil (Mora et al., 2011). In recent years, environmental metabarcoding has been proven to be able to fill this knowledge gap. This approach has enabled a direct access to microbial communities living in diverse

and unexplored habitats and thus, is being used to unveil the enormous diversity in eukaryotic tree of life. Needless to say that a unified approach based on traditional taxonomy, barcoding, and metabarcoding will expedite the identification of new species.

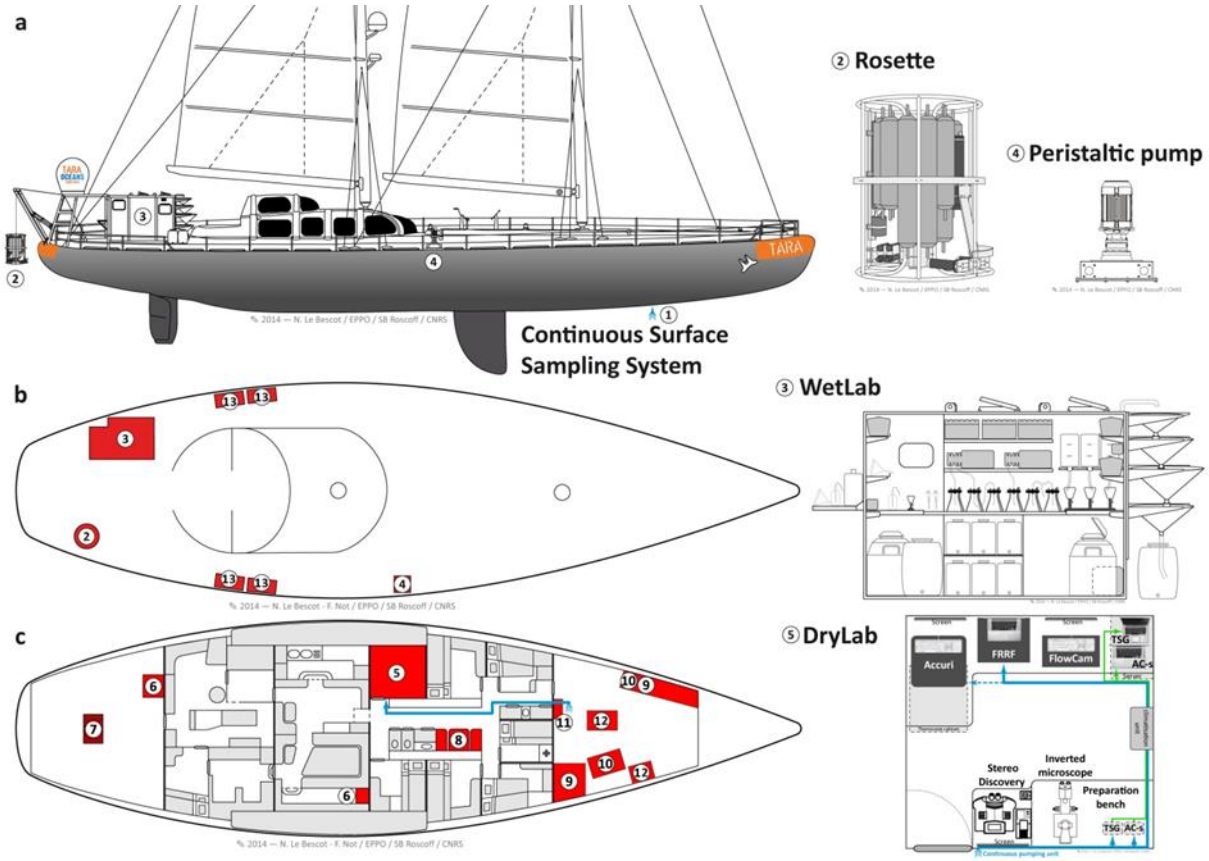


Figure 1.14. Sampling devices and working areas on-board SV *Tara* are shown from the vessel's (a) side-view, (b) bird's-eye-view of the deck, and (c) inside-view. They consist of the (1) continuous surface sampling system; (2) CTD-rosette; (3) wet lab; (4) peristaltic pump for large-volume sampling; (5) dry lab; (6) oceanography engineers working areas; (7) winch; (8) video imaging area; (9) storage areas at room temperature; (10) storage areas at +4°C and -20°C; (11) MilliQ water system (12) diving equipment, and (13) storage boxes. Water flow from the continuous surface sampling system to the dry lab is shown in blue. (Taken from Pesant et al., 2015).

1.4. Tara Oceans: a comprehensive sampling of marine planktonic biota

The name plankton, coined by German physiologist Viktor Hensen (1835-1924), is derived from the Greek word planktos, meaning “wanderer” or “drifter” (Thurman, 1997). Plankton are drifting organisms with limited powers of locomotion and are transported primarily by prevailing water movements. Plankton inhabit oceans, seas, lakes, and ponds, and their abundance and distribution strongly varies with light, nutrient availability, the physical state of the water column, and the abundance of other plankton. Plankton are primarily divided into three broad functional (or trophic level) groups, i.e. phytoplankton (autotrophic, prokaryotic or eukaryotic), zooplankton (small protozoans or metazoans that feed on other plankton), and bacterioplankton (bacteria and archaea). Plankton form the base of the marine food web and the variability in their population influences higher trophic levels. They encompass a wide diversity and help to regulate Earth’s climate. Plankton span several orders of magnitude in size and often exist in tight biotic interactions necessitating the development of an integrated, comprehensive, sampling strategy to best capture plankton community composition and ecology. The conventional, >150 year-old morphological view of marine eukaryotic plankton comprises ~11,400 catalogued species divided into three broad categories: ~5,700 species of metazooplankton (holoplanktonic animals), ~4,350 species of phytoplankton (microalgae), and ~1,320 species of protozooplankton (relatively large, often biomineralized, heterotrophic protists) (Sournia et al., 1991; Wiebe et al., 2010; Boltovskoy et al., 2005).

1.4.1. Background

Over many centuries global expeditions have led to major scientific breakthroughs, the H.M.S. Beagle (1831-1836) and the H.M.S. Challenger (1872-1876) voyages being notable examples. Ocean exploration now provides promising first steps towards understanding the role of the ocean in global biogeochemical cycles and the impact of global climate change on ocean processes and marine biodiversity. Recently, the Sorcerer II expeditions (2003-2010) (Gross, 2007) and the Malaspina expedition (2010-2011) (Laursen, 2011) carried out global surveys of prokaryotic metagenomes from the ocean’s surface and bathypelagic layer (>1000 m), respectively. The *Tara* Oceans Expedition (2009-2013) complemented these surveys by collecting a wide variety of planktonic organisms (viruses to fish larvae) along with extensive environmental data from the ocean’s surface (0-200 m) and mesopelagic (200-1000m) layers at a global scale. Moreover, *Tara* Oceans takes such surveys one step further by combining modern sequencing and state-of-the-art imaging technologies (Karsenti et al., 2011).

As a research infrastructure, the *Tara* Oceans Expedition mobilized over 100 scientists to sample the world oceans on board of a 36 m long schooner (SV *Tara*) refitted to operate state-of-the-art oceanogr-

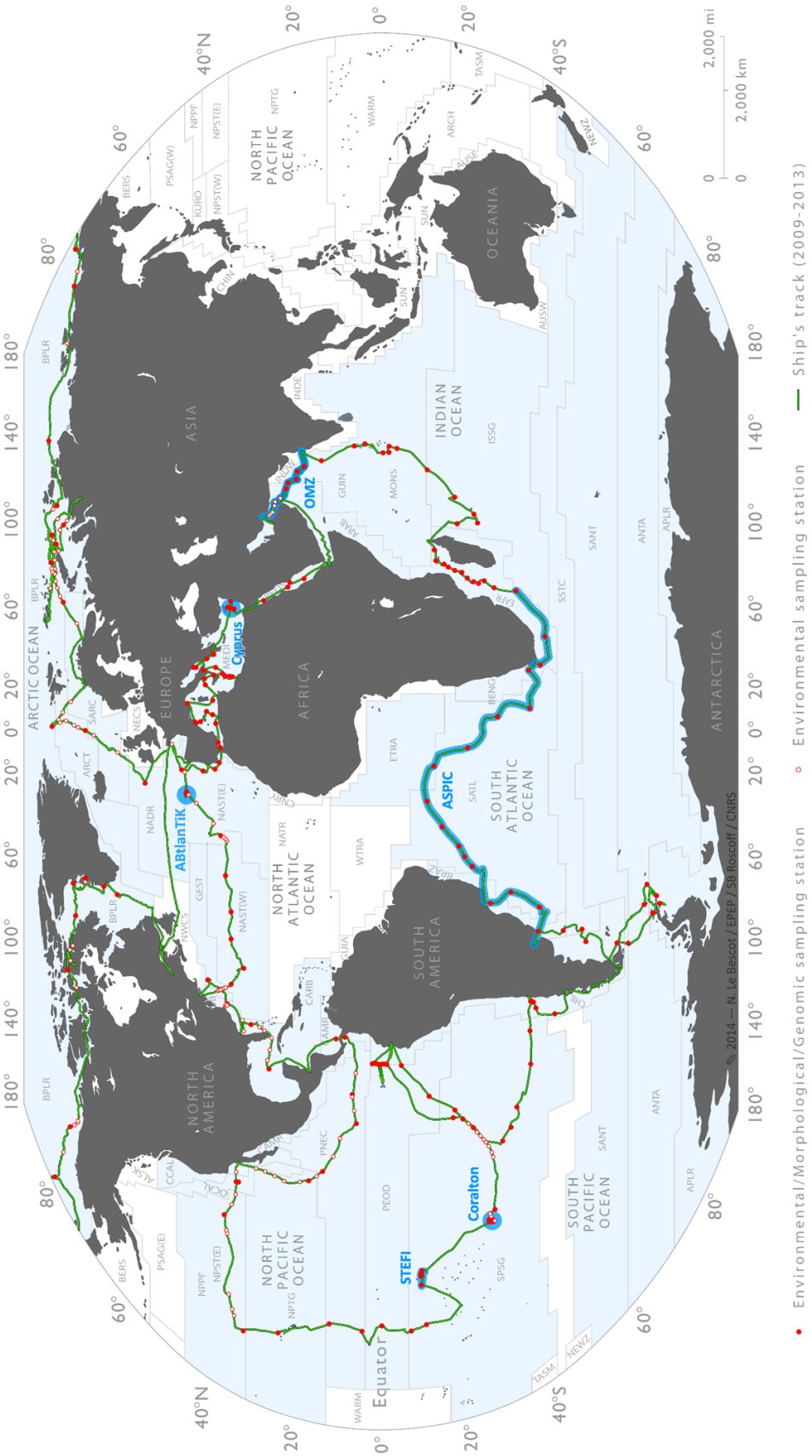


Figure 1.15. Sampling route of the *Tara* Oceans Expedition (green track), showing stations where plankton were sampled in their environmental context (full red dots) and where only environmental conditions were measured (open red dots). Sections of special scientific interest are identified along the sampling route (light blue). Longhurst's biogeographical provinces [Longhurst 2007] are shown in the background and those sampled during *Tara* Oceans Expedition are highlighted in blue. (Taken from Pesant et al., 2015).

-aphic equipment. On-board the schooner, the team was consistently composed of five sailorsand six scientists, including one chief scientist, two oceanography engineers in charge of deck operations, two biologists preparing and preserving samples for later morphological and genetic analyses, and one optics engineer in charge of imaging live samples on board. Working areas on-board (**Figure 1.14**) were setup to accommodate a smooth running of the operation. A winch was installed below the working deck, and sampling devices were deployed from the stern of the ship. A laboratory container (wet lab) equipped with water filtration systems was installed on the port side of the ship and an industrial peristaltic pump was installed on starboard to sample large volumes of water from various depths up to 60 m. A laboratory was set up inside the schooner (dry lab) at the center of the ship on port side, for imaging live organisms. The dry lab also contained flow through instruments connected to the underway continuous sampling system. The storage area was located in the forward hold.

Samples were stored on-board either in liquid nitrogen (three 36-L Dewar tanks), in a 400 L freezer (-20°C), in a 360 L fridge (+4°C), or at room temperature. Samples were packed and transported by World Courier (http://www.worldcourier.com/) in pre-conditioned temperature-controlled containers to ensure that the cold-chain was never interrupted. Frozen samples were kept in dry ice at all times.

1.4.2. Sampling strategy and methodology

The sampling strategy and methodology of the *Tara* Oceans Expedition (2009-2013) is presented as follows:

(a) Atmospheric and oceanographic context at mesoscale. The regular sampling programme was designed to study a variety of marine ecosystems and to target well-defined mesoscale features such as gyres, eddies, currents, upwellings or hot spots of high CO_2 (ocean acidification) or low oxygen (OMZ) concentrations (**Figure 1.15**). In order to identify these features before sampling, the atmospheric and oceanographic context were determined at the mesoscale using remote sensing products, arrays of Argo drifters and the meteorological station on-board Tara. Satellite observations (Chlorophyll a, sea surface temperature (SST), and altimetry) and real-time ocean model outputs (Mercator Ocean) were also used on a daily basis to revise sampling positions with respect to the selected oceanographic features. A total of 210 stations were characterized at the mesoscale to provide broader environmental context for the morphological and genomic study of plankton.

(b) Properties of seawater and particulate matter from physical, optical and imaging sensors mounted on the continuous surface water sampling system. Continuous measurements of surface water physical, chemical and biological properties were often used to fine tune the location of

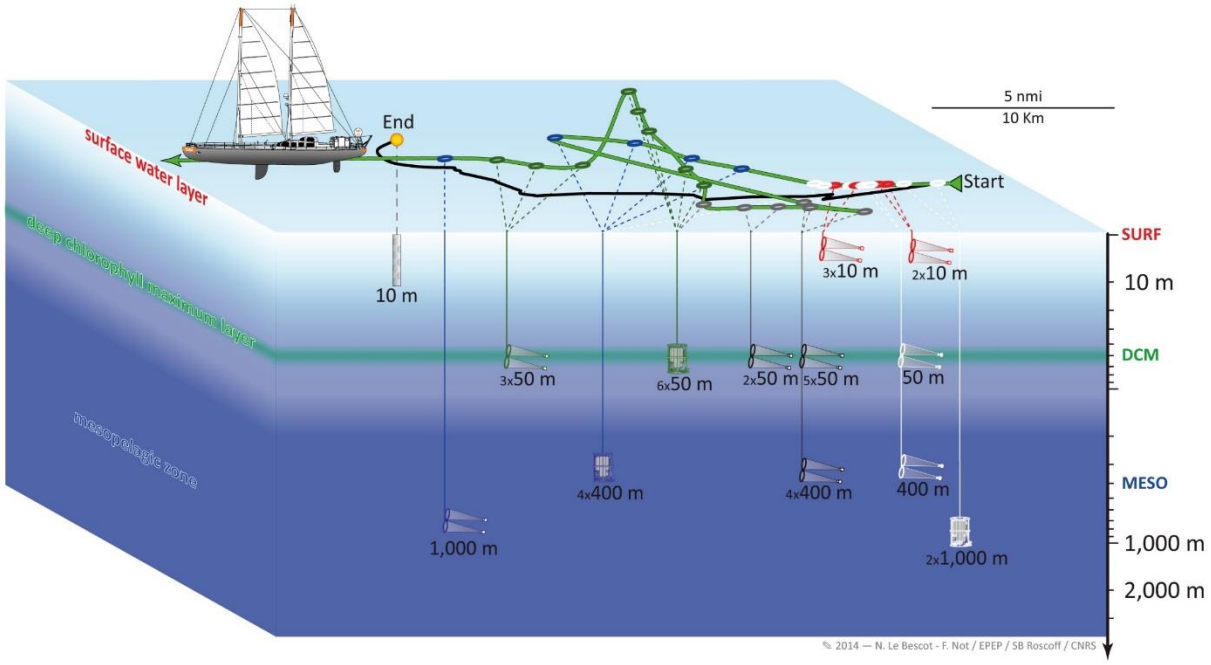


Figure 1.16. Geographical representation of the chronology (over 48 hours) and spatial distribution of sampling events during station TARA_152 in the North Atlantic Ocean (Start at -016.8454°E, 43.6850°N). At that station, an Argo drifter (10-m floating anchor and satellite positioning) was used to follow the water mass during sampling (black surface track). Along the route of SV *Tara* (yellow surface track), red, green, and blue markers correspond to sampling events targeting the "surface layer", "deep chlorophyll maximum layer", and "mesopelagic layer", respectively. White and grey markers correspond to the deployment of nets over a fixed depth layer, typically 0-100m or 0-500m during day and night time, respectively. (Taken from Pesant et al., 2015).

sampling stations that were initially selected based on satellite images. The in-line, continuous sampling system installed on SV *Tara* comprised a SeaBird TSG temperature and conductivity sensor, a WETLabs Ac-S spectrophotometer, a WETLabs chlorophyll fluorometer, and a Fast Repetition Rate Fluorometer (FRRF) to assess photosynthetic efficiency.

(c) Environmental features and sampling stations. During the *Tara* Oceans Expedition (2009-2013), plankton were collected from any of the three distinct oceanic features, i.e., "surface water layer", "deep chlorophyll maximum layer", and "mesopelagic zone" (**Figure 1.16**). The "surface water layer" (SRF) was simply defined as a layer between 3 and 7 m below the sea surface. The "deep chlorophyll maximum layer" (DCM) was determined from the chlorophyll fluorometer (WETLabs optical sensors) mounted on the Rosette Vertical Sampling System [RVSS-SENSORS]. The "mesopelagic zone" (MESO) corresponds to the layer between 200 and 1000 m depths. The sampling layer within the mesopelagic zone was selected based on vertical profiles of temperature, salinity, fluorescence, nutrients, oxygen, and particulate matter. The selected environmental feature varied from station to station. Some of these mesopelagic zone features have special scientific interest, such as the "oxygen minimum zone" (OMZ) and the "epipelagic mixing layer" (MIX).

All sampling devices used during the *Tara* Oceans Expedition (2009-2013), essentially were a High Volume Peristaltic pump [PUMP-SENSOR], a Rosette Vertical Sampling System [RVSS-SENSORS and NISKIN] and plankton nets [NET-TYPE-MESH]. Plankton were sampled in their environmental context at a total of 210 stations, of which 57 sampled only the surface layer, 62 sampled the surface layer and a second depth-specific feature, and 40 sampled the surface layer, the deep chlorophyll maximum layer and a third depth-specific feature.

(d) Properties of seawater and particulate & dissolved matter from physical, optical and imaging sensors mounted on the vertical profile sampling system. Repeated deployments of a Rosette Vertical Sampling System [RVSS] were essential to locate features that have a vertical component and have a signature below the surface, such as eddies, upwellings, fronts, deep chlorophyll maxima, and oxygen minimum zones. The [RVSS] was specifically designed for the *Tara* Oceans Expedition (2009-2013), using various SEABIRD® components [RVSS-SENSORS].

(e) Properties of seawater and particulate & dissolved matter from discrete water samples. In addition to sensors mounted on the Rosette Vertical Sampling System [RVSS], seawater was collected using Niskin bottles [RVSS-NISKIN] (6 x 8-L Niskins and 4 x 12-L Niskins) in order to further characterize a sampling station's environmental conditions. Measurements include pigment concentrations from

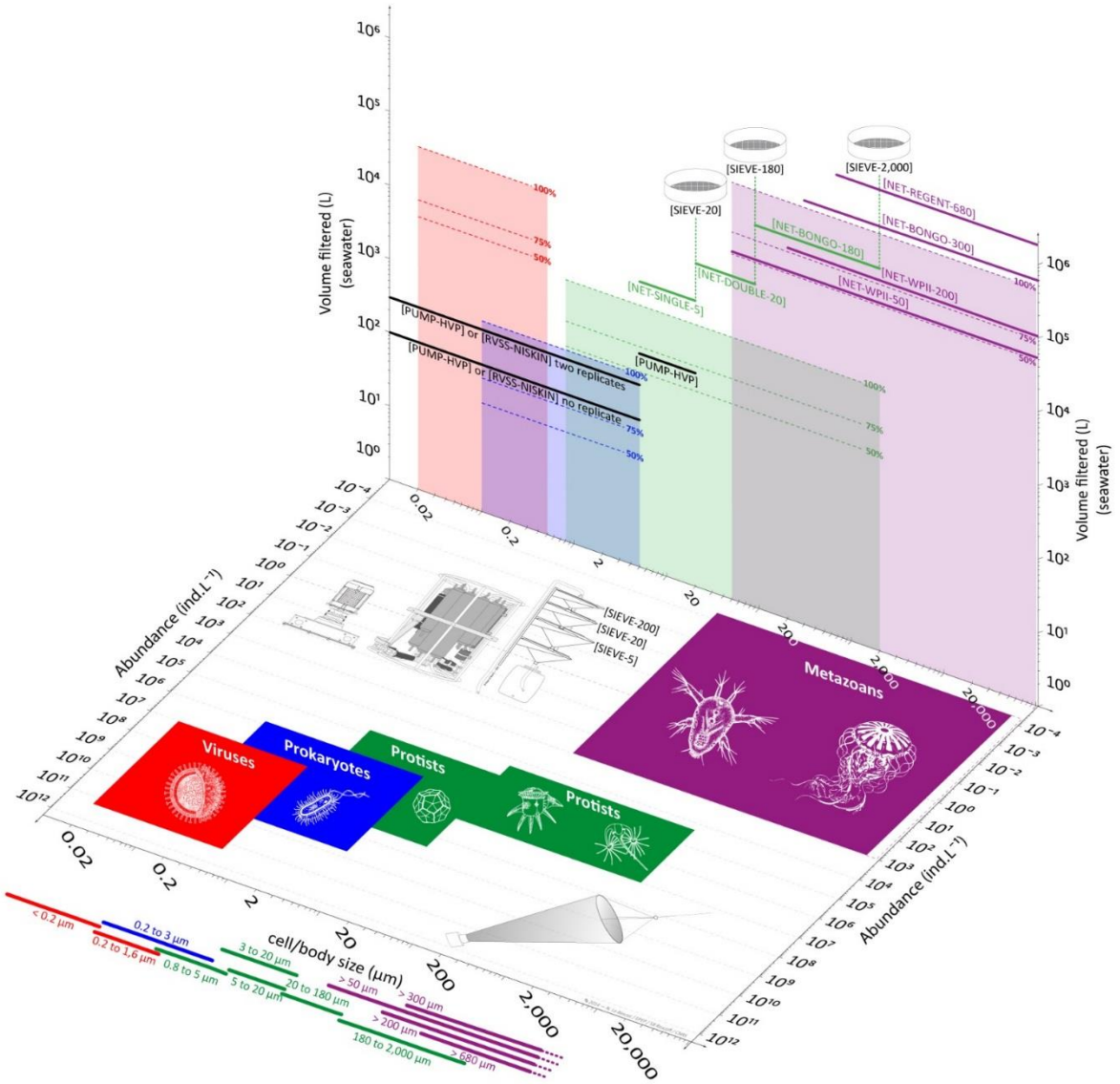


Figure 1.17. Empirical background for the *Tara* Oceans sampling methodology and the choice of sampling devices. The horizontal plane shows the range of body/cell size and natural abundances reported in the literature for viruses (including giant viruses), prokaryotes, protists and metazoans (coloured boxes). The sampling devices used to collect plankton <5 µm in size (i.e., high volume peristaltic pump and rosette with Niskin bottles) and >5 µm in size (i.e., plankton nets) are illustrated as well on the horizontal plane. The vertical plane shows the volume of seawater required to capture 100%, 75% and 50% of species richness reported in the literature for viruses (including giant viruses), prokaryotes, protists and metazoans (shaded boxes). The typical volume of seawater collected by sampling devices are shown in comparison (horizontal thick lines). Also illustrated on the vertical plane: sieves were used to remove large organisms in the case of plankton nets for protists (5, 20 and 180 µm mesh). (Taken from Pesant et al., in prep).

HPLC analysis (10 depths per vertical profile; 25 pigments per depth), the carbonate system (Surface and 400m; pHT, CO_2, pCO_2, fCO_2, HCO_3, CO_3, Total alkalinity, Total carbon, OmegaAragonite, OmegaCalcite, and dosage Flags), nutrients (10 depths per vertical profile; NO_2, PO_4, NO_2/NO_3, Si, quality Flags). More than 250 vertical profiles of these properties were made across the world ocean.

(f) Marine plankton. Plankton sampled during the *Tara* Oceans Expedition cover six orders of magnitude in size (10^{-2}-10^5 µm) and correspond to viruses, giant viruses (giruses), prokaryotes (bacteria and archaea), protists (unicellular eukaryotes), and metazoans (multicellular eukaryotes). These five groups form the bulk of biomass throughout the oceans and drive the global biogeochemical cycles that regulate the Earth system (Arrigo, 2005, Falkowski, 2008; Karl, 2007). Unicellular eukaryotes, or protists, cover a broad range of cell size (0.8-2000 µm). They are taxonomically very diverse with representatives in all of the 8 super-groups of the eukaryotic tree of life (Baldauf, 2005), whose roles in marine and Earth systems ecology are largely unexplored. Meso-zooplankton (metazoans; multicellular eukaryotes) range in size from 50 µm to several metres, and play a pivotal role in both the transfer of energy to higher trophic levels such as fish and other large predators, and in the vertical export of particulate matter produced at the surface of the ocean (Banse, 2013).

Given that abundance is generally an inverse function of cell size (**Figure 1.17**; horizontal plane), and that we are interested in capturing the diversity of both dominant and less abundant organisms, we used a series of sampling devices that collect/filter enough volume of seawater to capture the diversity of organisms in the following 10 size fractions: <5 µm (or <3 µm), 5-20 µm (or 3-20 µm), <20 µm, 20-180 µm and 180-2000 µm for sampling plankton viruses, prokaryotes and unicellular eukaryotes, and >50, >200, >300, >500 and >680 µm for sampling large plankton unicellular eukaryotes and metazoans. Whenever possible, replicate sampling was performed to assess plankton natural variability and to ensure long-term storage of samples in view of future re-analysis using new technologies, notably in the fields of high throughput imaging and -omics.

Sampling plankton viruses, prokaryotes and unicellular eukaryotes. Plankton sampling devices used to collect small size organisms (<20 µm size fractions) include Niskin bottles mounted on the rosette [RVSS-NISKIN] or occasionally attached individually on a line [LINE-NISKIN], a High Volume Peristaltic pump [PUMP-HVP], and exceptionally a 10-L plastic bucket [BUCKET]. The choice of the sampling device was determined by weather conditions and the depth of the targeted environmental features. The “surface water layer” was systematically sampled using the pump [PUMP-HVP], and exceptionally (at 3 stations) using a 10-L plastic bucket [BUCKET]. The “deep chlorophyll maximum layer” was sampled preferentially with the pump [PUMP-HVP] or alternatively using multiple deployments of the

rosette [RVSS-NISKIN] when the sampling depth was >60 m. The "mesopelagic layer" was systematically sampled using multiple deployments of the rosette [RVSS-NISKIN]. Nets [NET-SINGLE-5] were used for the 5-20 µm size-fraction. Plankton from the 20-180 µm size-fraction were collected using a double plankton net with a 20 µm mesh size [NET-DOUBLE-20]. Plankton from the 180-2000 µm size-fraction were collected using a 180 µm Bongo Net [NET-BONGO-180]. All the nets were lowered to the selected environmental feature and towed horizontally for 5-15 min. Upon recovery, all nets were rinsed from the outside with running seawater. After each use, nets, cod-ends, and sieves were rinsed with fresh water and checked for holes.

Sampling plankton large unicellular eukaryotes and metazoans. Plankton sampling devices used to concentrate and collect the larger and less abundant organisms (>50 µm size fractions) consisted of plankton nets with mesh sizes ranging from 50 to 680 µm [NET-TYPE-MESH] and metal pan-shaped sieves [SIEVE-MESH] to remove large organisms as needed. Upon recovery, all nets were rinsed from the outside with running seawater. After each use, nets, cod-ends, and sieves were rinsed with fresh water and checked for holes.

1.4.3. *Tara* Oceans integrated pipeline

High throughput imaging platform. The high-throughput imaging platform used by *Tara* Oceans includes (i) on-board and on-land flow cytometers to monitor virus particles, bacteria, and small protists, (ii) on-land digital and confocal microscopy for detailed 2D/3D imaging of cells within the 5–20-mm range, (iii) on-board and onland FlowCams and ZooScans for quantitative recognition of organisms ranging from 20 mm to a few cm, light sheet and confocal microscopes for 3D imaging, and (iv) on-land electron microscopes for detailed ultrastructural analyses of small protists and viruses (Karsenti et al., 2011).

High throughput sequencing methods. High throughput sequencing methods were used to obtain both deep phylogenetic rDNA/rRNA tag data (metabarcodes) and metagenomic and metatranscriptomic functional profiles from size fractions covering the entire plankton community from viruses to fish larvae (Karsenti et al., 2011).

Eco-, morpho- and genetic modelling. Tara Oceans aims to visualize, quantify, and genetically characterize ocean biodiversity within entire plankton ecosystems (Karsenti et al., 2011). The unprecedentedly comprehensive data sets are being employed to gain a deeper understanding of biodiversity gradients within and among systems and contrasting environments and can, thus, assist in establishing rules governing the self-organization of organism networks (Fuhrman, 2009; Raes et al.,

2011). Also, they can be used to develop predictions about how these rules and communities will be affected by a changing environment.

In summary, through a global network of researchers in more than 20 nations, the Tara Oceans Expedition (Karsenti et al., 2011) has engaged in a coordinated scientific program to develop the first planetary-scale data collection effort. Taken together, this comprehensive and systematic sampling strategy has opened the door to explore associations between biodiversity and function in marine planktonic ecosystem by integrating genomics, morphology and environmental data. This uniquely exhaustive expedition can be deemed as a much needed step towards integrating the biological complexity into predictive global-scale ecological models that can serve in managing the oceanic ecosystems in response to environmental changes. This integrated end-to-end interdisciplinary initiative will serve to address global issues from a holistic perspective.

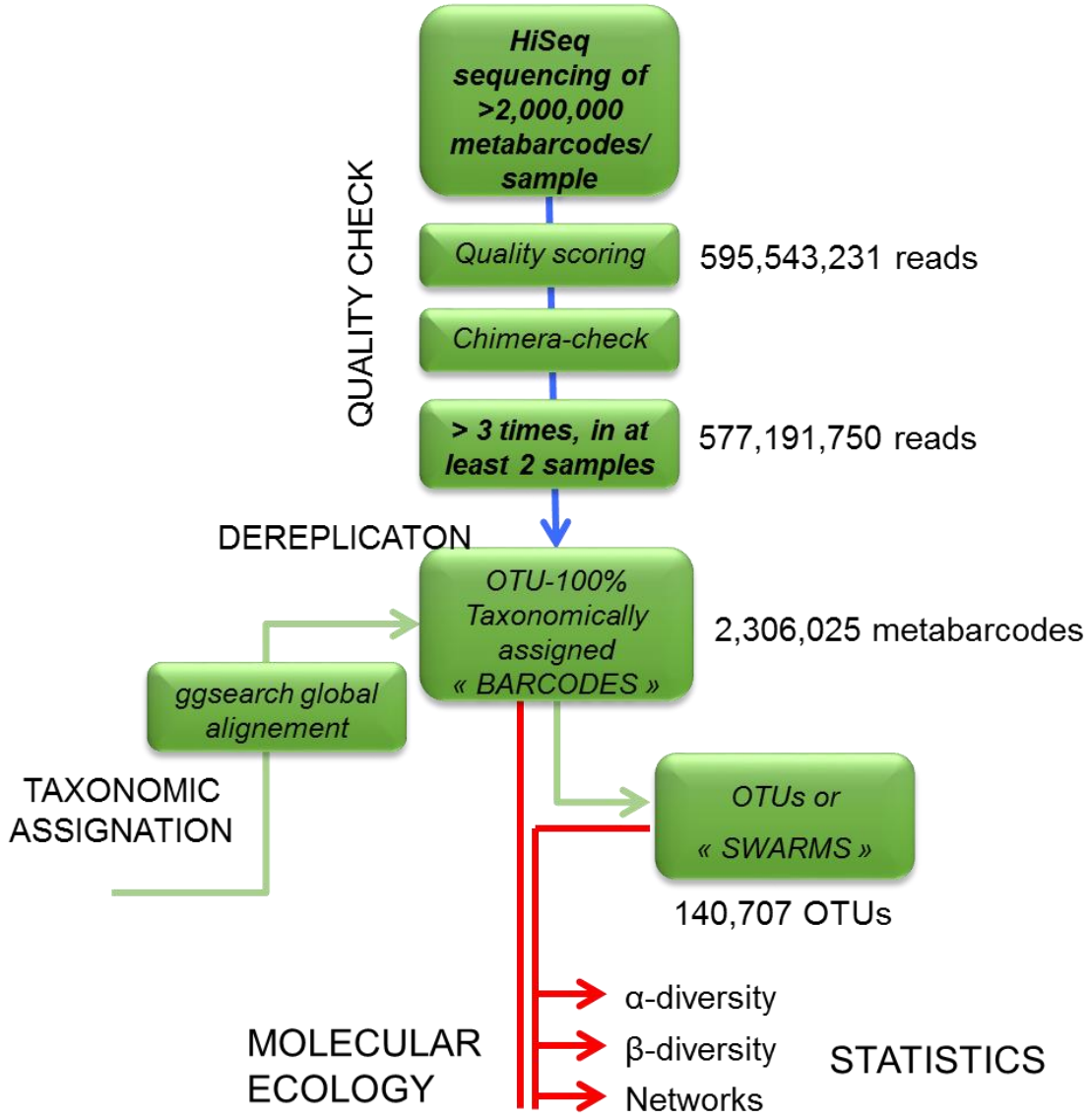


Figure 1.18. Bioinformatics pipeline: from raw Illumina rDNA read production to biodiversity and ecology analyses. Raw V9 rDNA reads were first filtered based on sequence quality scoring and chimera removal analyses, and only reads present in atleast 3 copies and 2 independent samples were considered for downstream analyses. Filtered reads were dereplicated and taxonomically assigned by homology (ggsearch global alignment) to an expert-curated database (V9_PR2). Metabarcodes -identical dereplicated reads- were finally clustered into OTUs (Operational Taxonomic Units) using the Swarm algorithm (Mahe et al. 2014) for α- and β-diversity analyses (de Vargas et al., 2015).

1.5. Aim of the thesis

The goal of my thesis was to assess the global diversity pattern of diatoms in world oceans and the impact of various abiotic factors on the diatom diversity. And these goals were accomplished by primarily looking into the following fundamental questions:

(a) How does the diatom abundance and diversity vary in the world ocean?

(b) Is the diversity consistent across different size classes and oceanic provinces?

(c) How are the diatom communities structured?

I used 63371 diatom metabarcodes (V9 region of SSU 18S rDNA) from 46 sampling sites sampled during *Tara* Oceans Expedition to address the following objectives:

(a) To develop and evaluate the metabarcoding approach for the assessment and analysis of diatom diversity, encompassing species richness as well as spatial heterogeneity.

(b) To investigate diversity choke points.

(c) To determine the variation in diatom diversity varies across different oceanic provinces and size classes.

(d) To test whether dispersal limitation affect diatom species richness, and how much variation in the community structure can be explained.

(e) To investigate the impact of spatial, environmental and biotic drivers on patterns of species richness and composition.

(f) To determine a substitute for classical diversity indices.

(g) To test how the commonness and rarity varies across different stations and size classes.

To explore holistic patterns of photic-zone eukaryotic plankton biodiversity, ~766 million ribosomal DNA (rDNA) sequence reads were generated from samples across the world's oceans collected during the *Tara* Oceans expedition (de Vargas et al., 2015). A global metabarcoding approach was designed to cover the majority of eukaryotic plankton diversity, encompassing four organismal size fractions covering the majority of: piconano-plankton (0.8-5µm), nano-plankton (5- 20µm), micro-plankton (20-180µm), and meso-plankton (180-2000µm). The V9 region of the nuclear 18S rDNA gene, that is suited for assessing general patterns of biodiversity of entire eukaryotic communities, was chosen as a barcode. A strict quality-check pipeline led to a final dataset of ~580 million reads or ~2.3 million unique sequence reads (**Figure 1.18**). These were assigned taxonomic entities by alignment to an expert-curated database (V9_PR2 database) containing 77,449 reference V9 rDNA barcodes representing 13,432 genera and 24,435 species from all known major lineages of the tree of eukaryotic life.

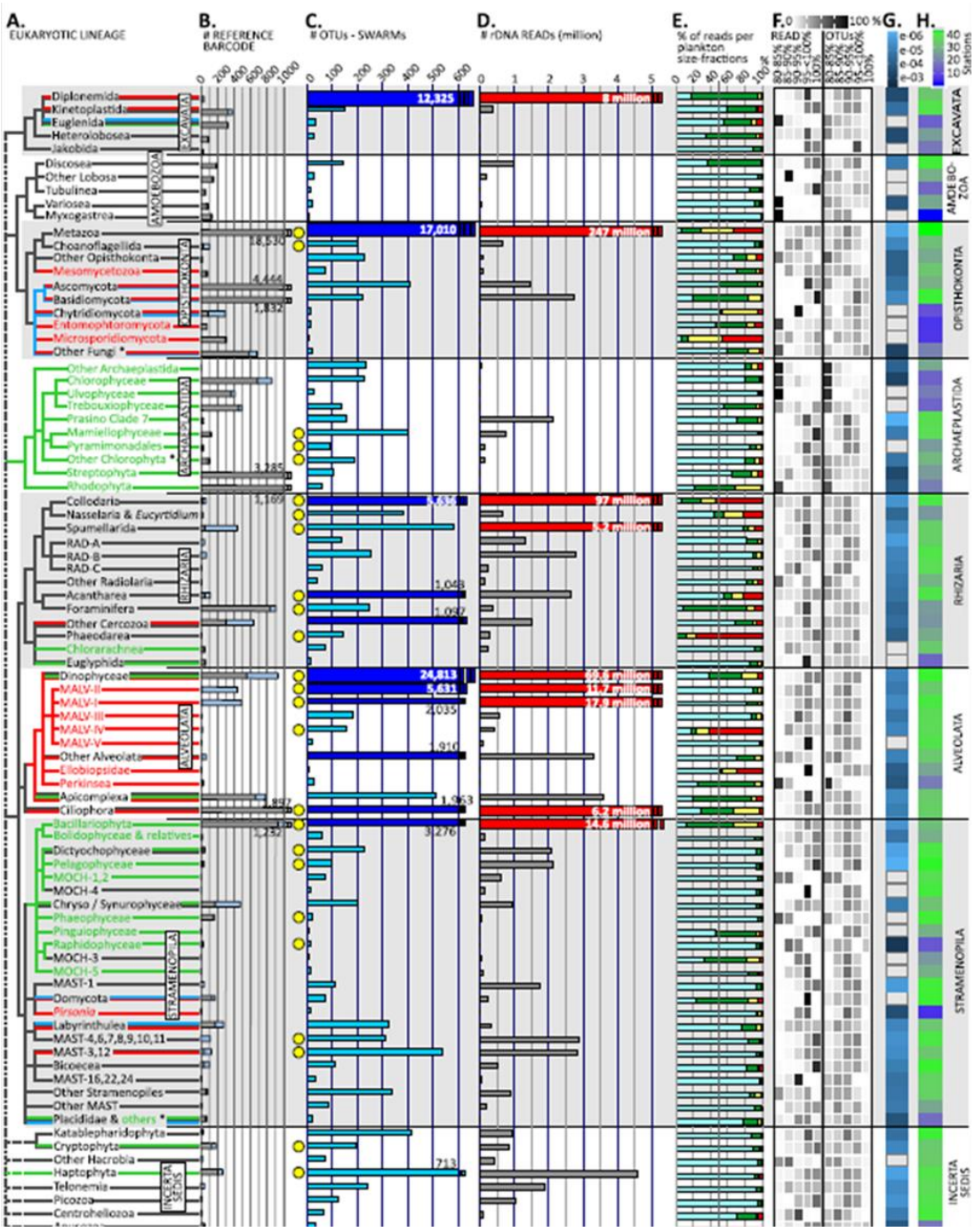


Figure 1.19. Phylogenetic distribution of the assignable part of eukaryotic plankton ribosomal diversity. (A) Schematic phylogeny of the 85 deep-branching eukaryotic morpho-lineages represented in our metabarcoding dataset, with broad ecological function (red = parasitic lineage; green = photoautotrophic lineage; blue = osmo/saprotrophic lineage; black = mostly hetero/phagotrophic lineage). Note that many well-known protistan lineages adapted specifically to marine benthic or terrestrial habitats were totally missing from photic-zone plankton and thus do not appear in the tree. **(B)** Number of reference V9 rDNA barcodes used to annotate the metabarcoding dataset (grey = with known taxonomy at the genus and/or species level; light blue = from previous 18SrDNA environmental clone libraries). **(C)** Tara-Oceans V9 rDNA OTU richness (the dark-blue thicker bars indicate the 11 hyper-diverse lineages containing >1,000 OTUs). **(D)** Eukaryotic plankton abundance expressed as numbers of rDNA reads (the red bars indicate the 9 most abundant lineages with >5 million reads). E. Proportion of rDNA reads per organismal size fraction (light blue = piconano-; green = nano-; yellow = micro-; red = mesoplankton). **(F)** Percentage of reads and OTUs with [80-85%], [85-90%], [90-95%], [95-<100%], [100%] sequence similarity to a reference sequence. **(G)** Slope of OTUs rarefaction curves. **(H)** Mean geographic occupancy (de Vargas et al., 2015).

Sequencing of close to 2 million V9 rDNA reads from each of the 334 size-fractionated plankton samples was sufficient to approach saturation of both local and global eukaryotic biodiversity. This survey unveiled ~75% of eukaryotic ribosomal diversity in the photic zone of the ocean. The extrapolated total richness of ~150,000 OTUs is much higher than the ~11,400 formally described species of marine eukaryotic plankton, and given the relatively low taxonomic resolution power of V9 rDNA barcodes, it likely represents a highly conservative, lower boundary estimate of the true number of eukaryotic species in this ecosystem.

About one third of ribosomal diversity in *Tara* Ocean's dataset did not match (at ≥80% identity) any sequence in the extensive V9_PR2 reference database. This significant unassignable diversity did not represent a large component (2.6% of reads), and increased in both richness and abundance in smaller organismal size fractions, suggesting that it corresponds to rare and minute taxa that have escaped traditional morphology-based analyses. The remaining 87,000 OTUs could be classified amongst 103 monophyletic groups covering the full spectrum of catalogued eukaryotic diversity, i.e., the 7 recognized super-groups and multiple incertae sedis lineages whose origins go back to the primary radiation of eukaryotic life in the Neo-Proterozoic. Many well-known lineages adapted specifically to marine benthic or terrestrial habitats were totally missing from photic-zone plankton and thus do not appear in **Figure 1.19**. The conventional, >150 year-old morphological view of marine eukaryotic plankton comprises ~11,400 catalogued species divided into three broad categories: ~5,700 species of metazooplankton (holoplanktonic animals), ~4,350 species of phytoplankton (microalgae), and ~1,320 species of protozooplankton (relatively large, often biomineralized, heterotrophic protists) (Sournia, et al., 1991; Wiebe et al., 2010; Boltovskoy et al., 2005). A largely congruent picture of the distribution of diversity between and within these organismal categories exhibited a typically 3 to 8 times more rDNA OTUs than described morphospecies in the best-known lineages within these categories. This is within the range of the number of cryptic species typically detected in many globally-distributed pelagic taxa (e.g., de Vargas et al., 1999; Halbert et al., 2013). Despite this overall congruency for classical categories, OTUs related to morphologically-described taxa represented only a minor part of total eukaryotic plankton phylogenetic diversity. Only 31 of the 87 (36%) lineages represented in **Figure 1.19** have been regularly recognized in previous plankton biodiversity studies, 11 of which were known almost exclusively from clone library surveys (**Figure 1.19C**).

Overall, less than 1% of OTUs were strictly identical to reference sequences, and OTUs were on average only 86% similar to any reference sequence (**Figure 1.19G**), indicating that the large majority of photic-zone eukaryotic plankton diversity had not previously been sequenced from cultured strains, single-cell isolates, or rDNA clone library surveys. Assessment, on a group by group basis, of the amount of

phylogenetic information added to current knowledge of total protistan rDNA diversity by *Tara* Oceans metabarcodes indicated that mean tree length increase was of 453%, reaching >100% in 43 lineages. Even in the best-referenced groups such as the diatoms (1,232 reference sequences, **Figure 1.19B**), abundant novel rDNA sequences were recorded both within known groups and forming entirely new clades. Surprisingly, >80% of eukaryotic plankton biodiversity was found in heterotrophic lineages. 11 'hyper-diverse' lineages contained >1,000 OTUs, together representing ~88% and ~90% of all OTUs and reads, respectively (**Figure 1.19C**), amongst which the only phototrophic taxa were diatoms and about a third of dinoflagellates, together comprising ~20% and ~18% of 'hyper-diverse' OTUs and reads, respectively.

Most hyper-diverse photic-zone plankton belonged to three super-groups, the Alveolata, Rhizaria, and Excavata, all poorly known in terms of their biology and functional ecology. The Alveolata, mainly consisting of phagotrophic and parasitic taxa (ciliates and most dinoflagellates, MALVs and apicomplexans, respectively), were by far the most diverse super-group, with ~42% of all assignable OTUs. The Rhizaria are a group of amoeboid heterotrophic protists with active pseudopods that also display a wide spectrum of ecological behavior from phagotrophy to parasitism and symbioses sensu lato (Burki and Keeling, 2014). Rhizarian diversity peaked in the Retaria, the group including giant protists that build complex skeletons in silicate (Polycystinea), strontium sulfate (Acantharea), or carbonate (Foraminifera), and comprise key microfossils for paleoceanography. Enormous unsuspected rDNA diversity (5,636 OTUs) was recorded within the Collodaria, which are mostly colonial, poorly silicified or naked polycystines that typically live in obligatory symbiosis with photosynthetic dinoflagellates and display remarkably complex behaviors (Swanberg, 1974). Arguably the most surprising component of novel biodiversity was the >12,300 OTUs related to reference sequences of diplonemids, an excavate lineage that has only two described genera of flagellate grazers, one of which parasitizes diatoms (Schnepf, 1994). Their ribosomal diversity was much higher than that observed in classical plankton groups such as Foraminifera, ciliates, or diatoms (50-fold, 6-fold, and 3.8-fold higher, respectively), and was furthermore one of the only phylo-groups far from saturating (**Figure 1.19E**).

Beyond these hyper-diverse eukaryotic lineages, our dataset revealed considerable phylogenetic diversity (>50 deep-branching groups) of unknown or very poorly known phagotrophic, osmotrophic, and parasitic protists. These results fundamentally challenge the common view of plankton diversity, inspired from terrestrial ecology, whereby phytoplankton and metazooplankton make up ~90% of eukaryotic plankton diversity and heterotrophic protists correspond to a minor compartment reduced in food web modeling to a single entity, often idealized as ciliate grazers.

This data set demonstrates that the diversity of eukaryotic plankton in the photic-zone of the world's ocean is significantly higher than previously thought, but that it is a finite compartment whose taxonomic, functional, and ecological properties can be addressed holistically using a functional metabarcoding approach. In years to come, decoding the ecological and evolutionary rules governing this exceptional diversity will be essential for understanding one the most critical biomes for the functioning of the Earth system.

1.6. Thesis outline

The results from this study are organized in three results chapters (**Chapters 2-5**) followed by conclusions and perspectives (**Chapter 6**) beyond this general introduction (**Chapter 1**).

Chapter 2 provides a comprehensive understanding of global biodiversity patterns and structure of planktonic diatom communities across the world ocean by analysing photic zone samples from 46 stations of the *Tara* Oceans sampling project (2009-2013) (Karsenti et al., 2011). The analysis was based on the use of environmental ribosomal DNA fragments ("meta-barcodes") as markers of diatom biodiversity. The Protist Ribosomal Reference database (PR2; Guillou et al., 2013) was used to assign a traditional taxonomy to meta-barcodes, thus linking classical knowledge of diatom biodiversity to an analysis based on high-throughput sequencing. I then explored whether general patterns in the structure of diatom biodiversity emerge across size classes, genera and ecological niches. The abundance of different genera of diatoms varies primarily between size classes (pico-, nano-, micro- and meso- plankton), independently of the ecosystem studied and the sampling period.

Chapter 3 assesses the impact of environmental (niche-based approach) and spatial (neutral processes) drivers in explaining the differences in species richness and composition pattern. In this chapter, the effects of various abiotic and biotic environmental variables, along with spatial variables were assessed for each sub-community (based on size) using multiple regression and canonical redundancy analysis (RDA) and their partial form to control for spatial variables effects. The observed distribution pattern of diatom barcode assemblages in the world ocean suggests that connectivity of local water masses to ocean circulation has a major impact on marine diatom biogeography.

Chapter 4 investigated if the co-occurring ribotypes exhibit a distinct behavior and response to environmental conditions in a way that they can be expressed as a function of varying environmental parameters. This study aims to identify the distributional pattern for the identified clusters co-occurring ribotypes and to examine the relative importance of environmental factors in explaining the structure of each cluster.

Chapter 5 presents an in-depth analysis to explore the patterns of species abundance by employing rank abundance distributions (RADs) along with commonness and rarity patterns of protists in the world's ocean. Rank abundance curve was obtained for each sample and its shape, especially of the tail, was studied. The plotted RADs for all the samples under study showed a heavy-tailed distribution which appears to follow a power-law behavior. A framework was developed to calculate and classify

the exponent of the tail. The underlying objective of this work was to relate the characteristic shapes and slope of the tail of RAD curve to some feature of the environment.

In **Chapter 6**, I have discussed the importance of the results in identifying interconnections between associated theories and underlying drivers, and deduced implications for future studies. Promising novel research questions and directions are identified to explore how these marine communities are structured and to determine core community assembly rules. In the context of this thesis, I propose that the meta-barcoding approach provides a potential framework to investigate environmental diversity at a global scale, which is deemed as an essential step in answering a wide range of ecological research questions.

CHAPTER 2

Insights into the Biogeographical Patterns of Planktonic Diatom Diversity – an Assessment Using Metabarcoding

Summary

Manuscript to be submitted to PNAS Plus

Unexpected High Diversity of Diatom Communities in the Open Ocean

Authors

Shruti Malviya[1], Eleonora Scalco[2], Stéphane Audic[3], Alaguraj Veluchamy[1,$], Lucie Bittner[1,†], Flora Vincent[1], Daniele Iudicone[2], Colomban de Vargas[3], Eric Karsenti[4,1], Adriana Zingone[2], Chris Bowler[1],*

Affiliations

[1]Ecology and Evolutionary Biology Section, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR8197 INSERM U1024, 75230 Paris, France

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

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

[4]Directors' Research, European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany

[†]Current address: Université Pierre et Marie Curie Paris 06, Équipe Analyse de données à haut débit en génomique, Département des Plateformes, Institut de Biologie Paris Seine (IBPS/FR 3631), 9, quai Saint- Bernard, 75005 Paris, France.

[$]Current address: CDA, KAUST, Jeddah

Abstract

Diatoms (Bacillariophyta) constitute one of the most diverse and ecologically important phytoplanktonic groups. They are considered to be particularly important in nutrient-rich coastal ecosystems and at high latitudes, but considerably less so in the oligotrophic open ocean. The Tara Oceans circumnavigation collected samples from a wide range of oceanic regions using a standardized sampling procedure, allowing a broad comparison of plankton composition. In the present study, a total of approximately 12 million diatom V9-18S rDNA tags has been analyzed to explore diatom global diversity and community composition. This was done using 293 size-fractionated planktonic communites collected at 46 sampling sites across the global ocean's euphotic zone. Based on the total assigned ribotypes, Chaetoceros was the most abundant and diverse genus, followed by Thalassiosira, Corethron and Fragilariopsis. We found only a few cosmopolitan ribotypes, displaying an even distribution across stations and high abundance, many of which could not be assigned to a known genus. Three distinct clusters from Equatorial, Mediterranean and Southern Ocean waters were identified that share a substantial percentage (~25-42%) of ribotypes within them. Sudden drops in diversity were observed at Cape Agulhas, which separates the Indian and Atlantic Oceans, and across the Drake Passage between the Atlantic and Southern Oceans, indicating the importance of these ocean circulation choke points in constraining diatom diversity patterns. We also observed unexpectedly high diatom diversity in the open ocean, suggesting that diatoms may be more relevant in these oceanic systems than generally considered.

Keywords: Biodiversity, diatoms, metabarcoding, *Tara* Oceans, choke points, open ocean

2.1. Introduction

Diatoms are single-celled photosynthetic eukaryotes deemed to be of global significance in biogeochemical cycles and the functioning of aquatic food webs (Smetacek, 1998; Falkowski, 2002; Armbrust, 2009). They constitute a large component of aquatic biomass, particularly during conspicuous seasonal phytoplankton blooms, and have been estimated to contribute as much as 20% of the total primary production on Earth (Nelson et al., 1995; Field et al., 1998; Falkowski et al., 1998). They are widely distributed in almost all aquatic habitats, except the hottest and most hypersaline environments, and can also occur as endosymbionts in dinoflagellates and foraminifers (Round et al., 1990). However, planktonic diatoms prefer cold, nutrient-rich regions encompassing polar as well as upwelling and coastal areas (Crosta et al., 2005).

Diatoms constitute one of the most diverse planktonic groups with a wide range of species estimates (ranging between 10,000 and 200,000), including between 1,400 and 1,800 marine planktonic species

(Sournia et al., 1991). An early report on the number of diatom species on Earth, based on extrapolation and compilation of morphological data, estimated that there exist more than 200,000 species (Mann and Droop, 1996). More recently, Guiry (2012) suggested a 'conservative figure of 12,000 described species of diatoms, with a further 8,000 to be discovered'. However, in recent years molecular studies have demonstrated the presence of cryptic diversity in many diatom taxa (e.g. Lundholm et al., 2006; Behnke et al., 2004), which also corresponds to mating incompatibility (Amato et al. 2005) and considerable functional diversity (Degerlund et al., 2012; Huseby et al., 2012). In view of that, a recent overview (Mann and Vanormelingen, 2013) estimated that the total number of extant species is at least 30,000.

Despite their wide distribution, Vanormelingen et al. (2008) suggested that cosmopolitan diatom species are not the rule and that a considerable degree of endemism is likely in diatom communities. A few years later, Mann and Vanormelingen (2013) proposed an "intermediate dispersal hypothesis" stating that "long-distance dispersal is rare, but not extremely rare." The quantification of diatom diversity and its variation across space (and time) is therefore important for understanding fundamental questions of diatom speciation.

Characterization of diatom diversity requires accurate and consistent taxon identification. However, given the apparently large numbers of undescribed species, together with the cryptic and semi-cryptic diatom species, morphological analyses alone fail to provide a complete description of diatom diversity. Therefore, genetic investigations on environmental samples can be effectively used to consistently quantify and understand diversity for more robust and rapid community comparison (Rodríguez-Martinez et al., 2013). The past decade has seen tremendous advances in the field of molecular and biochemical methods for rapid taxon identification and characterization. The introduction of DNA sequence data to systematics has facilitated the discovery of numerous previously undescribed taxa, revealing distinct species identified by subtle or no morphological variations (e.g., Beszteri et al., 2007; Sarno et al., 2005). Allozyme electrophoresis (Gallagher 1980; Soudek and Robinson, 1983), DNA fingerprinting (Rynearson and Armbrust 2000), isozyme analysis (Skov et al., 1997), and microsatellite marker analysis (Evans and Hayes, 2004) have also been used to assess diatom diversity at lower (intraspecific) taxonomic levels.

Several efforts have been undertaken to develop alternative approaches for diversity analysis using sequence-based approaches on environmental DNA (Steele and Pires, 2011; Bik et al., 2012). With the advent of high-throughput DNA sequencing, DNA metabarcoding has now emerged as a rapid and effective method to develop a global inventory of biodiversity that cannot be detected using classical

microscopic methods (Yu et al., 2012; Ji et al., 2013; Sogin et al., 2006; Meusnier et al., 2008; Bittner et al., 2013). Metabarcoding combines DNA-based identification and high-throughput DNA sequencing and is based on the premise that differences in a diagnostic DNA fragment coincide with the biological separation of species. Limitations have been identified for metabarcoding (Will et al., 2005; Coissac et al., 2012; Bittner et al., 2013), mainly by its dependency on PCR and thus exposure to amplification artefacts (Bellemain et al., 2010; Taberlet et al., 2012), by its susceptibility to DNA sequencing errors (Coissac et al., 2012), and by the considerable investment required to build comprehensive taxonomic reference libraries (Taberlet et al., 2012; Yoccoz, 2012). However, compared to other classic methods metabarcode data sets are far more comprehensive, many times quicker to produce, and relatively independent from taxonomic expertise.

Metabarcoding is hence receiving considerable interest, especially when classical species identification is time consuming and difficult due to the presence of diverse and widespread cryptic species, or when the organisms are too small to be observed. However, the choice of variable DNA region to be barcoded needs to be evaluated carefully (Riaz et al., 2011). For eukaryotes, recent reports have proposed the use of partial 18S rDNA sequences as potential molecular markers (Amaral-Zettler et al., 2009). The 18S rDNA contains nine hypervariable regions (V1-V9) (Ki and Han, 2005). Amaral-Zettler et al. (2009) first employed the V9 region to assess general patterns in protistan diversity. They suggested that these regions have the potential to assist in uncovering novel diversity in microbial eukaryotes.

Owing to their complex evolutionary history, diatoms have a “mix-and-match-genome” (Armbrust, 2009) which provides them with a potential range of abilities like the presence of proton-pump-like rhodopsins (Slamovits et al., 2011), ice-binding proteins (Janech et al., 2006), biogenic silica formation (Kroger and Sumper, 2000), and a urea cycle (Allen et al., 2011). These abilities probably underlie their success in occupying a wide range of ecological niches. Despite the key role of diatoms in the functioning of many ecosystems, their biodiversity and geographical distributions are poorly understood at a global scale. Most of the research in this area has been focused towards understanding the patterns of biological diversity in a particular diatom genus of interest at a local or regional scale (e.g., Nanjappa et al., 2014). Here, we have performed a global analysis of diatom community composition using the V9 region of 18S rDNA as diversity marker. To achieve this, we employed taxonomic profiling of 293 samples derived from forty-six sampling sites (de Vargas et al., 2015) along the Tara Oceans circumnavigation (Karsenti et al., 2011). Experimental validation of the molecular data was established by light microscopy using samples from selected sites. Our study provides significant and novel insights into the current patterns of diatom genetic diversity for the first time on a global scale.

2.2. Results

Our study, summarized in Figure 1, was structured to develop a framework for a rapid molecular-based analysis of diversity. The results are presented in three broad sections, namely (i) evaluation of the V9 hypervariable region as a diversity marker for diatoms, (ii) correlation between the molecular and morphological estimates, and (iii) global biogeographical patterns exhibited by diatoms.

2.2.1. Evaluation of V9 region of 18S rDNA as a diversity marker for diatoms

To evaluate the limits of metabarcoding in a given organismal group, it is important to assess how the variable region under study relates to the genetic diversity in discriminating various taxa (Taberlet et al., 2012; Ficetola et al., 2010). In the present study, 2947 full-length 18S rDNA sequences were obtained from the PR2 reference database corresponding to 718 diatom species (Guillou et al., 2013). They were aligned and entropy along the full length was computed. The sequence variations along the entire length was used to assess the nine hypervariable regions (V1-V9) using the *RNAstructure* program (**Fig. S1A**). Regression of V1-V9 p-distances by Neighbor-Joining (NJ) algorithm onto those of full length 18S sequences showed that a combination of hypervariable sites can help in better discrimination of different species of diatoms. The performance of V9, the sequence used for *Tara* Oceans metabarcoding (de Vargas et al., 2015), was 23% less than that of the full-length 18S sequence, and taxa assignment at less than 70% identity in the V9 region was found to be insufficient for diatoms (**Fig. S1B)**. Although the mean genetic distance for V9 was higher than the full length 18S rDNA, the phylogeny based on that fragment was found to be less reliable as compared to the full length sequence. Length variation and pairwise genetic distances calculated using the Kimura-2-parameter model for all nine hypervariable regions are shown in **Figures S1B and S1C**. We found that the V4 region wrongly placed some raphid pennate diatoms within centric groups, whereas the V9 region could not differentiate well between radial and polar centric diatoms, nor between raphid and araphid pennate groups (**Fig. S1D**). Further, the resolving power of the V9 region was evaluated by computing pairwise p-distance between and among different genera. The results showed an average intergenus p-distance of 0.134, which is about two times larger than the mean intragenus distance (0.065), indicating fairly good discrimination ability exhibited by V9 at the genus level.

2.2.2. Global dataset of diatom V9 metabarcodes

A total of ~580 million quality-checked reads, representing ~2.3 million unique rDNA ribotypes (V9 region of 18S rDNA), were generated from 334 photic-zone plankton communities sampled during the *Tara* Oceans expedition (de Vargas et al., 2015). Taxonomy assignments for all ribotypes were obtained through annotation against an expert-curated V9 reference database (for details, see de Vargas et al.,

2015) using the global alignment search strategy implemented in the ggsearch36 program (*Fasta* package). This reference database contains sequences from both cultured strains and the environment, and contained 1,232 unique diatom V9 reference sequences corresponding to 159 genera, with most genera being represented by more than one sequence (**Fig. S2**). Of the 159 genera in the reference dataset, we retrieved 86 genera in our dataset. However, only 76 out of 86 were assigned at an identity greater than 85% and were selected for further analysis.

For the present study, 293 global samples encomapssing 46 stations from the photic zone (sub-surface (SRF) and deep chlorophyll maximum (DCM)) were used that corresponded to four size classes (0.8-5 µm, 5-20 µm, 20-180 µm, 180-2000 µm). A total of 65,404 V9 rDNA diatom-assigned ribotypes (represented by ~14.6 million reads) were retrieved from the 293 communities. Rarefaction analysis indicated that the 65,404 diatom ribotypes approached saturation at a global scale (**Fig. 2A**), although individual oceanic regions such as the North Atlantic Ocean (NAO) and Red Sea (RS) were far from saturation. Preston log-normal distribution extrapolated the true diatom ribotype richness to 96,710 ribotypes, of which 33,339 represent Preston's veil and thus remain undiscovered. This data suggests that our survey has retrieved ~67% of diatom ribosomal diversity in the photic zone of the global ocean (**Fig. 2B**). Using the 'swarm' approach (Mahé et al. 2014), all the ribotypes were clustered into biologically meaningful operational taxonomic units (OTUs), yielding 4416 distinct OTUs. Each OTU was represented by the most abundant ribotype in the 'swarm' cluster. For these swarms, Preston's veil revealed the completion in sampling to be 81% with an extrapolated number of OTUs to be 5468 (**Fig. S3**). The total number of OTUs was found to be two to three times the number of diatom species recognized in the marine plankton (1400-1800 species) in the literature (Sournia et al., 1991).

Diatoms were found to be one of the most represented eukaryotic phototrophic lineages (#2 in eukaryotic phototrophic lineages and #5 with respect to all marine eukaryotic lineages) (de Vargas et al., 2015). Overall, diatom reads accounted for about 2.86% of the total eukaryotic reads retrieved in our set of samples, but represented more than 25% of the total eukaryotes at some locations, e.g., in the Southern Ocean (SO) (**Fig. 2C**). They formed 4.86% of the protist community; Collodaria (Radiolaria) being the most prominent protistan lineage (39% protistan reads). Diatoms contributed ~75% to the total photosynthetic community at Station 11 (MS, dominated by *Leptocylindrus*), more than 78% and 65% at Stations 84 and 85, respectively (SO), 44% at Station 82 (SO), and more than 38% and 44% at Stations 122 and 123, respectively (Marquesas Islands; SPO), and globally represented 27.7% of the total eukaryotic photosynthetic planktonic community. The mean percentage of diatom reads across 46 stations were 2.6%, 5% and 19.9% with respect to the total eukaryotic reads, protistan reads and photosynthetic reads, respectively (**Fig. 2C**). Stations in the MS (Stations 18, 20 and 30), RS (Stations

31, 32 and 33), Indian Ocean (IO; Stations 41, 45 and 48), South Atlantic (SAO; Stations 72, 76, 78) and SPO waters (Station 98) were found to be very scarce in diatoms in comparison to other photosynthetic groups such as dinoflagellates and haptophytes. In general, the normalized abundance of diatoms showed a significant decrease from coastal to open ocean (e.g., from Stations 65-67 to 68-78).

2.3.3. Diatom community composition

Nearly 55% of the reads corresponded to ribotypes (33,327) assigned at least up to the genus level and the large majority (>90%) of these assigned sequences belonged to planktonic genera. Of the 76 genera found, *Chaetoceros* was found to be the most abundant and diverse genus, representing 23.4% of total assigned sequences. *Thalassiosira* accounted for 14.4% of total assigned sequences, followed by *Corethron* (11.5%), *Fragilariopsis* (11.5%), *Leptocylindrus* (10.4%), *Actinocyclus* (8.9%), *Pseudo-nitzschia* (4.6%) and *Proboscia* (4.1%) (**Fig. 3a**). However, there were also few sequences that were assigned to genera known from freshwater or benthic environment, but in many cases with a quite low similarity (e.g. *Fragilariforma* and *Epithemia*). The MARine Ecosystem biomass DATa (MAREDAT) project previously provided global abundance and biomass data for all major planktonic diatoms of the global ocean ecosystem (Leblanc et al., 2012). Our dataset showed an overlap of 44 diatom genera with MAREDAT (**Fig. S4**), while 32 diatom genera from our study were not found in MAREDAT, indicating the comprehensiveness of our dataset. A total of 23 genera present in both MAREDAT and the reference database were not found in our dataset. Most of the unmapped genera were either freshwater (e.g., *Fragilariforma*, *Tabellaria*, *Ulnaria*, *Urosolenia*) or benthic and marine littoral species (e.g., *Amphiprora, Caloneis*, *Ardissonea*, *Hyalodiscus*, *Pseudostriatella, Entomoneis, Phaeodactylum*) except for only a few pelagic marine genera (e.g., *Bacterosira*, *Shionodiscus*). Some of these have only been reported in northern latitudes, which may explain their absence in our data set.

Intragenus diversity was found to vary from as low as one ribotype per genus (e.g., *Nanofrustulum, Asteroplanus, Bellerochea, Tenuicylindrus*) to as high as 6287 ribotypes (*Chaetoceros*) (**Fig. 3b**). *Chaetoceros* and *Thalassiosira* also accounted for the highest number of OTUs (**Fig. 3c**). The 5-20 and 20-180 µm fractions contained the highest numbers of diatom ribotypes, as expected, although an unexpectedly high number were also found in the smaller size fractions, derived from smaller species (e.g., *Nanofrustulum*, *Cyclotella*, *Minutocellus* and *Minidiscus*) or probably broken cells of larger species (e,g, *Attheya*, *Ditylum*, *Bellerochea, Licmophora*). The 180-2000 µm size fraction contained the lowest number of ribotypes, mostly from chain-forming diatoms (e.g., *Hyalosira*, *Fragilaria*), epizoic species (*Pseudohimantidium*), but also from small cells probably ingested by larger organisms, or retained in that fraction due to net clogging (e.g., *Nanofrustulum*). A clear distinction was seen in the distribution among different size-fractions, e.g., *Minidiscus, Epithemia, Licmophora, Attheya* and *Minutocellus*

were found highly restricted to the smallest size fractions on the one hand whereas genera like *Asterionellopsis, Lauderia* and *Odontella* were found principally in 20-180 µm size-fractions (**Fig. 3d**). *Pseudohimantidium* was found principally in the largest size fraction, consistent with it being known to attach to marine copepods of the genera *Corycaeus, Euterpina* and *Farranula* (Skovgaard and Saiz, 2006; Fernandes and Calixto-Feres, 2012; Garate-Lizarraga and Muneton-Gomez, 2009). Although some stations showed a majority of ribotypes in SRF samples (e.g., Stations 65, 67 and 85), others contained significantly more ribotypes in the DCM (e.g., Stations 52 and 66). Different genera were also found to prefer different depths, such as *Actinoptychus*, *Corethron*, *Coscinodiscus*, *Fragilariopsis*, *Leptocylindrus* and *Rhizosolenia* in sub-surface samples while *Asterionellopsis, Bellerochea, Helicotheca*, *Nanofrustulum* and *Lithodesmium* were seen mostly in DCM samples (**Fig. 3e**). The level of percentage identity to the reference sequence also varied across genera (**Fig. 3f**). At a level of 85% identity, a total of 63,371 ribotypes could be assigned to diatoms, although almost half of them (31,178) could not be assigned to any genus and were placed in one of the five unassigned classes mentioned above. *Pseudo-nitzschia*, *Actinocyclus*, *Attheya*, *Chaetoceros*, *Eucampia*, *Fragilariopsis, Minutocellus* and *Thalassiosira* were among the most cosmopolitan genera (see below), whereas many others (e.g., *Leptocylindrus)* were restricted to only a couple of stations (**Fig. 3g**).

2.2.4. Unassigned sequences/ Novelty

Overall, the sequences that could not be unambiguously assigned to any diatom genus on the basis of V9 rDNA annotation represented between 30% and 82% of diatom communities at different sampling sites. In general, unassigned ribotypes were particularly common in SPO and IO, with almost similar percentages at both depths (**Fig. 4A**). The diatoms in the smallest size fraction contributed most to the unknown sequences, with depth having no significant impact (**Fig. 4B**). The numbers of unassigned sequences in different oceanic provinces was generally consistent with the intensity of diatom research previously conducted in those areas, with MS and SO containing the best characterized diatom communities (**Fig. 4C**). On the other hand, the larger size-fractions (20-180 µm and 180-2000 µm) contained the highest numbers of assigned ribotypes, again mainly in MS and SO, consistent with microplanktonic diatoms being the most common and the best studied. The highest numbers of unassigned diatom ribotypes from all the size-fractions are from SPO and RS (>65%).

2.2.5. Comparison between light microscopy and V9 ribotype counts

To investigate whether V9-based relative abundance estimates for diatoms are comparable to community composition studies based on classical morphological identification methods using light microscopy (LM), diatom counts from 20-180 µm fractions were compared between the two methods

for eleven sampling stations. A simple comparison was initially disappointing, however the correlation between the two kinds of data was significantly improved when "unassigned" and "not known" sequences were deleted from the V9 dataset, and when some specific adjustments were applied (see Methods) (**Fig. 5**). A few cases of mismatch still persisted, e.g., the surface sample from Station 84 was dominated only by *Fragilariopsis* sp. in LM counts while *Chaetoceros* and *Fragilariopsis* were equally dominant genera along with unknown centric diatoms in the V9 dataset. However, the overall match between the two data sets was sufficiently close, thus indicating that V9 counts can provide a reliable estimate of diatom relative abundance in a given sample.

2.3.6. Global diversity patterns

We next examined intragenus diversity (expressed as effective number of species; ENS) and distribution in different oceanic contexts for the twenty most abundant genera. *Pseudo-nitzschia*, *Chaetoceros* and *Thalassiosira* were the most diverse genera whereas *Corethron*, *Leptocylindrus*, *Minidiscus* and *Planktoniella* were among the least diverse (**Fig. 6A**). Most diatom genera were seen in most oceanic provinces, e.g., *Actinocyclus*, *Eucampia, Proboscia* and *Pseudo-nitzschia.* Their abundance pattern was highly variable across provinces, for instance, *Chaetoceros*, *Corethron* and *Fragilariopsis* were highly abundant in SO, in accordance with previous data (e.g., Gersonde and Zielinski, 2000); *Attheya*, *Planktoniella* and *Haslea* were seen principally in SPO (**Fig. 6A**); *Leptocylindrus* was found to be highly abundant in MS, especially at Station 11, in line with reports from the Gulf of Naples (Ribera d'Alcalà et al., 2004) and other Mediterranean sites (Siokou-Frangou et al., 2010). In terms of global biogeography, the diversity of each genus (expressed as the number of ribotypes) was found to be strikingly variable across the oceans (**Figs. 6B and S6**).

Among surface samples, diversity and evenness across oceanic provinces varied greatly, attaining the highest values in RS, while among the DCM samples IO showed the highest diversity; SO was the least diverse at both depths (**Fig. 7A**). In terms of richness, the SO stations consistently showed the highest values owing to the presence of a majority of very low abundant ribotypes. Considerable variation in terms of overall ribotype diversity in different size fractions was observed (**Fig. S7**). In contrast with what was observed globally for marine planktonic eukaryotes in the *Tara* Oceans data set (de Vargas et al., 2015), diatom diversity did not consistently decrease with increasing size (**Fig. S7**). There were also no discernable differences in diatom diversity patterns between SRF and DCM samples.

Diatom diversity followed a latitudinal gradient, albeit weak (**Fig. 7B**). However, a comparatively stronger gradient was seen in the 20-180 size fraction in contrast to there being almost no gradient observable in the largest size fraction. A sudden drop in diversity was also observed in the Agulhas

retroflection region going from IO (Station 65) to SAO (Stations 66/67/68), and across the Drake Passage from SAO (Station 78) to SO (Stations 82/84/85) (**Figs. 7C**). Diversity was significantly lower in the samples from the Maldives garbage zone (Station 45, North Indian Ocean), but increased towards the north and the south (**Fig. S7**).

2.2.7. Community similarity

Diatom-annotated ribotype distribution patterns were generally consistent across all the stations in that only a few ribotypes were abundant and the large majority of the richness was attributed to the rare ribotypes (Malviya et al., manuscript in preparation). The number of different ribotypes per station varied from as low as 46 (Station 48) to as high as 16,100 (Station 85), with a mean richness of 4927. In general, it was found that the more abundant a ribotype, the more ubiquitous it was distributed **(Fig. 8A)**. Several ribotypes with considerable abundance but low occupancy were also seen, possibly indicating endemism but more likely due to the marked seasonality in diatom occurrence. It was found that "rare" ribotypes constitute a substantial fraction and at the same time they tend to be different at different sampling sites (Malviya et al., manuscript in preparation). Only 23 ribotypes were found in ≥90% of the studied sites, although these represented nearly 24% of the total relative abundance. The majority of these cosmopolitan ribotypes could not be assigned to a known diatom taxon (**Fig. 8A**).

The total number of ribotypes seen in MS, RS, IO, SAO, SO and SPO were 13119, 4586, 23722, 16269, 26846 and 29203, respectively. To first assess biogeography, we observed that most of the ribotypes in SO (53.3%), SPO (33.7%) and MS (26.9%) were not found elsewhere, whereas only a few ribotypes were specific to RS (2.3%). Similarly, IO (14.2%) and SAO (10.4%) showed only a small number of ribotypes specific to them (**Fig. 8B**). Altogether, nearly 52% (32,850 out of 63,371) of the ribotypes were seen only in one province. Interestingly, most of the ribotypes were shared between a combination of two provinces (in particular, SPO and IO (12,176 ribotypes), SAO and SPO (9,501 ribotypes), SAO/IO (8,569 ribotypes), and SO/IO (7,330 ribotypes)), whereas only 576 ribotypes (out of 63,371; 0.9%) were present in all oceanic provinces (**Fig. 8B**).

We then assessed similarity among surface stations for which all four size fractions were available (37 stations). Stations in SPO, SO, and MS showed the highest degree of internal similarity (**Fig. 9**). The clustering of stations revealed four major groups, including one for each of these three provinces, and one containing stations where diatoms were only present at low abundance. Each of these clusters shared a considerable percentage (~20-37%) of ribotypes within them. The community in MS was most distinct from the others, while IO showed the most similarity with the others (**Fig. S9A**). Non-metric

multidimensional scaling (NMDS) indicated that communities grouped according to oceanic provinces, albeit with significant overlapping (stress=0.19) (**Figs. S9B and S10**). The SPO and MS stations were nonetheless each seen to cluster together without any overlapping and the SO stations showed a very distinct community structure. Each of these oceanic clusters were significantly different (ANOSIM; R = 0.58; W = 0.001).

2.3. Discussion

In this study we explored diatom distribution and diversity employing a short hypervariable region (V9, ~130 base pairs) of the small subunit of the 18S rDNA gene as a diversity marker. The availability of a taxonomically comprehensive reference database, highly conserved primer binding sites, and the potential of V9 to explore a broad range of eukaryotic diversity makes this sequence well suited as a biodiversity marker (de Vargas et al., 2015). We are conscious that its resolution should be evaluated for each organismal group under study. For the diatoms, we show here that while it may have limited resolution at the species level, it is nonetheless well suited to explore genus level diversity. Another potential caveat of metabarcoding is the presence of multiple copies of SSU rDNA in some species with respect to others (Godhe et al., 2008; Galluzzi et al., 2004; de Vargas et al., 2015), which may challenge the use of rDNA barcodes for diatom diversity analysis. The diversity estimates obtained in this study should therefore be interpreted conservatively, as ribosomal diversity rather than species diversity. Nonetheless, we argue that our diversity data are congruent, as demonstrated by the match between molecular and morphological methods. The overall coherence between these two methods indicates that, at least within the diatoms, the number of rDNA copies per genome are generally comparable, in contrast to dinoflagellates (Godhe et al., 2008). A further limitation is that our data set is based on a single sampling event at each location, whereas there is known to exist substantial temporal variation in community structure (Nolte et al., 2010). Despite this, it is noteworthy that by using the metabarcoding approach we can compare richness/diversities among areas and within areas because undersampling and underrecording are less important issues as compared to morphological approaches. Moreover, the extent of the dataset undoubtedly allows an unprecedented look at diatom community structure on a global scale.

All the sampled communities followed comparable structural patterns, characterized by a few dominant ribotypes representing the majority of abundance and a large number of rare ribotypes constituting a long tail of rare sequences. A high number of v9 reads (~1.6 million) assigned to *Chaetoceros* indicated it to be the most dominant diatom genus consistent with previous morphological surveys, followed by *Thalassiosira, Corethron, Fragilariopsis, Leptocylindrus,*

Actinocyclus (~0.5-1 million).Our results showed that the top ten genera together accounted for more than 92.4% of the assigned reads. The dominance of these genera in the world oceans is similar to findings from other studies (e.g., Hinder et al., 2012). Regarding the biogeography of these most represented genera, our results clearly suggest that despite being widely distributed, all dominant genera do not exhibit similar abundance and diversity patterns across stations. Among the top ten genera, *Leptocylindrus* and *Attheya* displayed distinct geographical preferences, i.e., MS and SPO, respectively. Interestingly, it was observed that *Chaetoceros*, *Corethron* and *Fragilariopsis* were more abundant in SO, in agreement with previously reported data (Smol and Stoermer, 2010), whereas *Thalassiosira*, *Actinocyclus*, *Pseudo-nitzschia*, *Proboscia*, and *Eucampia* showed almost even world-wide distributions across all provinces (e.g., Chamnansinp et al., 2013). These results are in agreement with evidence indicating that most diatom genera are likely to be cosmopolitan due to a high chance of large scale dispersal (Vanormelingan et al., 2008). Notably, except *Navicula* and *Pleurosigma*, some genera, like *Skeletonema, Nitzschia, Achnanthes* and *Cocconeis*, which are known to be common/abundant in coastal waters were under-represented in our dataset. For a sampling site, the maximum median ribotype diversity was observed for *Chaetoceros, Thalassiosira* and *Pseudo-nitzschia*, whereas the minimum was seen in *Minidiscus, Planktoniella* and *Leptocylindrus*. However, the diversity within each genus varied greatly across stations suggesting variations in community structure, which warrant a more detailed analysis of the factors/processes influencing the distribution and diversity of each genus.

Fourtanier and Kociolek (2003) have catalogued 900 diatom genera whereas our reference database has only 159 genera, indicating that many genera lack sequence information. Indeed, nearly 50% of the ribotypes remain unassigned due to the lack of representatives in the reference database. It is noteworthy that one third of the diatoms represented in the MAREDAT database do not have ribotype assignments. Moreover, some genera are represented by only one reference sequence which may also affect the assignation of some sequences. This also explains the assignment of some sequences to freshwater or benthic genera in our dataset. Previous studies have estimated nearly 1,400-1,800 marine planktonic species (Sournia et al., 1991), whereas our results estimate 5,468 marine diatom planktonic species. To our knowledge, this is the largest dataset that allows to assess the total number of diatom species. Together with this, there is nonetheless likely to be a considerable amount of novel diversity within the diatoms as we found a higher proportion of unassigned ribotypes in areas that have historically been undersampled, such as the South Pacific and Indian Oceans. As shown in **Figure 8A**, we have several abundant and cosmopolitan ribotypes that remain unassigned due to the lack of suitable reference sequence. Their future identification will thus lead to an enormous increase in the assignable fraction of diatoms. To explore the identity of these novel ribotypes, there is an ongoing

attempt to clone and sequence larger portions of the corresponding rDNA gene (results not shown). In addition to these attempts, building an exhaustive and representative reference databases represents the most critical issue limiting sequence assignation. In the future, continued efforts in this direction will lead us towards a more complete quantification of novelty and diversity.

In general, marine planktonic diatoms are associated with nutrient rich waters and high biomass that are commonly found in coastal waters, upwelling areas or during seasonal blooms in the open oceans, such as the North Atlantic spring bloom (Cervato and Burckle, 2003; Bopp et al., 2005; Armbrust, 2009). Although our dataset only contains a few coastal sampling sites, the results reported here confirm that diatoms constitute a major component of phytoplankton and are most common in regions of high productivity (upwelling zones) and high latitudes (Southern Ocean). Furthermore, we show that diatom diversity is also high in offshore oligotrophic areas. At these sites, while diatom abundance is low (likely because their growth is limited most of the time), they are able to survive and be ready to take advantage of favorable ecological conditions when they arise. This reservoir of diversity is likely an essential asset ensuring an overall plasticity of response of the whole diatom community to environmental variability.

Across the Indian Ocean, diversity decreases irregularly towards southwest stations from a high diversity epicenter in the Red Sea. Within the open ocean stations, diatom abundance was more uniform and significantly lower than those in the Southern Ocean (Stations 84 and 85) and in upwelling zones (Stations 67 and 82), characterized by low water temperature and high nutrient concentrations. Our study identified two diversity choke points, between Stations 65 and 67, and 78 and 82. These stations were situated at different sides of the Agulhas retroflection and the Drake Passage, respectively. Both areas are known to be choke points for ocean circulation (Siedler et al., 2013; Cunningham et al., 2003). We also observed that diversity tended to be higher in open ocean stations in comparison with nearby coastal stations. Previous studies on diatom fossil records reported that the Agulhas choke point is not a barrier to plankton dispersal (Cermeño and Falkowski, 2009). However, a recent study (Villar et al., 2015) reported strong contrasts in richness across the choke point and suggested that Agulhas rings are a means of connectivity between the basins. The second choke point is constrained by the Antarctic Circumpolar Current (ACC), and is an important conduit for exchange between the Atlantic, Southern, and Pacific Oceans. At the Drake Passage, the Antarctic Circumpolar Current branches off to give rise to the Malvinas Current which flows northward over the Argentine slope and outer shelf transporting saline, cold, nutrient-enriched waters (Peterson and Stramma, 1991). The high abundance of diatoms at Station 82 can be attributed to these nutrient-enriched waters being transported by the Malvinas Current.

A more detailed analysis of community similarity showed that the spatially separated and isolated sampling sites clustered together, suggesting that these communities have evolved so that they are extremely similar, supporting the concept of convergent evolution of communities. A closer analysis displayed an interesting biogeographical pattern via clustering stations based on their latitudes. The five major clusters identified represent five distinct latitudinal bands suggesting a strong spatial impact on community structure. Previous studies have reported negative latitudinal gradients as a common pattern in marine ecosystems (Fuhrman et al., 2008; Sul et al., 2013). However, the extent to which these gradients can be generalized to marine communities remains uncertain. Diatom diversity also exhibited these large-scale diversity gradients, albeit only weakly. At the same time we understand that these regional and latitudinal patterns can also be results of non-synoptic sampling, meaning that we sampled distant areas in different periods/seasons of the year. Also, we cannot ignore the bias due to a single sampling event at each station.

Our results indicate that diatoms exhibit wide geographical ranges, with a low to moderate structuring consistent with the oceanic provinces sampled. However, it was difficult to interpret any strong evidence in favor of local communities being structured under similar assembly rules across the world. A substantial number of ribotypes were seen exclusively in Pacific and Southern Ocean waters, which may suggest endemism. Also, various studies have reported that the Southern Ocean has high biodiversity and high endemism owing to its longer period of geographic isolation (Dayton, 1994). As mentioned above, endemism can also be attributed to the bias because of single sampling. But, here we would like to emphasize the unprecedented depth of sequencing which should have potentially covered everything.

Based on the data reported here, Baas Becking's hypothesis that "everything is everywhere, but the environment selects" (Baas Becking, 1934) holds only partially for diatom distributions. The worldwide distribution of different ribotypes from the most abundant diatom genera suggest that these protists have evolved to adapt to varying environmental conditions to exploit a range of ecological niches. This can be thought of as the underlying cause of ecotype differentiation that has made diatoms the most successful group of protists. The current study, using a metabarcoding approach, demonstrates the environmental diversity prevalent in highly adaptable phytoplanktonic diatom communities at a global scale. It has addressed a more generalized question on how marine communities are structured by assessing ecological patterns in their distribution and diversity. This study has thus laid a foundation for investigation in the direction of understanding the processes involved in structuring marine diatom communities and controlling their biodiversity. Along with this, the various physico-chemical

parameters and other contextual data collected during the *Tara* Oceans expedition has an immense potential to address a range of ecological questions raised by this study.

2.4. Materials and methods

2.4.1. Distance based Analysis

The PR2 database v99 (Guillou et al., 2013) contains 2947 full length 18S unique diatom sequences. These sequences were aligned and sequence variations along the entire sequence were used to define the hypervariable regions. Entropy calculation was done on all reference sequences. Pairwise distances were calculated for the full length and all hypervariable regions using Kimura-2-parameter model (Tamura et al., 2013). V4 and V9 sequences were used to check the performance in differentiating the four prominent phylogenetic clades of diatoms, i.e., radial centric, polar centric, araphid pennate and raphid pennate. Each of the V4 and V9 hypervariable regions and full-length 18S rDNA sequences were aligned using MUSCLE and phylogenetic inference was done with NJ algorithm using pairwise distances in MEGA5. The tree was statistically tested using 1000 bootstraps.

A reference database was obtained and all the reference sequences were aligned. Shorter sequences (less than 125 nucleotides) along with extremities were eliminated to obtain same sequence lengths. To evaluate the ability of V9 region to differentiate between the intragenus and intergenus variation among diatom V9 sequences, we calculated p-distance between all pairs of reference sequences.

2.4.2. Metabarcoding dataset

The *Tara* Oceans expedition (Karsenti et al., 2011) collected 293 planktonic samples from 46 sampling stations, from seven oceanographic provinces, i.e. North Atlantic Ocean (NAO), Mediterranean Sea (MS), Red Sea (RS), Indian Ocean (IO), South Atlantic Ocean (SAO), Southern Ocean (SO), and South Pacific Ocean (SPO). At each station, plankton communities were obtained for four size fractions from two water-column depths (SRF and DCM). Total nucleic acids (DNA + RNA) were extracted from all samples, and the hyper-variable V9 region of the nuclear 18S rDNA was PCR-amplified (Amaral-Zettler et al., 2009). The V9 reads were quality checked and to reduce the influence of PCR and sequencing errors, only sequences seen in at least two different samples with at least 3 copies were retained, giving a total of ~580 million reads represented by ~2.3 million unique metabarcodes (de Vargas et al., 2015). These unique barcodes were taxonomically assigned to known eukaryotic entities based on the PR2 database (Guillou et al., 2013). From this, metabarcodes assigned to diatoms, at a percentage identity of ≥ 85% to the reference sequence, were selected. All the barcodes were clustered into biologically meaningful operational taxonomic units (OTUs) using the 'Swarm' approach described by

Mahé et al. (2014). This method uses 1 base pair difference (local threshold) between barcodes. It also overcomes input-order dependency induced by centroid selection, a typical bias of classical clustering methods (Mahe et al., 2014). Various environmental variables, such as temperature, salinity, chlorophyll, oxygen, nitrate, phosphate and silicate, were recorded for each sample.

2.4.3. Morphological analyses

The samples selected for microscopy analyses included SRF and DCM samples from the Cape Agulhas region (st52, st64, st65, st66, st67, st68), the South Atlantic transect (st70, st72, st76, st78), the Antarctic stations (st82, 84 and 85), and south Pacific stations (st122, st123, st124, st125) (for full details of the sampling protocols used during the *Tara* Oceans expedition, see Pesant et al., 2015). Three ml of each sample was placed in an Utermöhl chamber with a drop of calcofluor dye (1:100,000), which stains cellulose thus allowing to better detect and identify diatom species. Cells falling in 2 or 4 transects of the chamber were identified and enumerated. Phytoplankton species were identified and enumerated using a light inverted microscopy (Carl Zeiss Axiophot200) at 400x magnification. The identification was performed at the species level when possible.

2.4.4. Taxonomy-based clustering

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

2.4.5. Global distribution analysis

Deviations from Preston's log-normal distribution was used to estimate the completeness of richness sampled. Also, the information from the samples was used to extrapolate the number of ribotypes that might be found if sampling is more intensive. The relation between abundance, occurrence and evenness of each ribotype was assessed. Pielou's evenness (Pielou, 1966) and exponentiated Shannon-Weiner H' diversity index (Hill, 1973), were used as an estimate of diversity. The percentage of shared ribotypes were calculated for each pair of stations and spearman correlation was used as a distance measure to cluster stations. Compositional similarity between stations were computed based on Hellinger-transformed abundance matrix and incidence matrix using Bray-Curtis and Jaccard indices respectively, as a measure of beta diversity. Non-metric multidimensional scaling (NMDS) was performed to visualize the level of similarity between different stations. The analysis of similarities

(ANOSIM) was used to test whether the groups were significantly different. For all statistical analyses, a value of $P < 0.05$ was considered significant. All the data analyses were performed in R (v.2.14.1).

Figure legends

Figure 1. Flow Diagram showing the material and method used in the study.

(A) Global diversity analysis was carried out using samples drawn from 46 global stations. At each station, eukaryotic plankton community was sampled at two depths (sub-surface (SRF) and Deep chlorophyll Maximum (DCM)), and fractionated into four size classes (i.e. 0.8-5 µm, 5-20 µm, 20-180 µm and 180-2000 µm), corresponding to 293 samples altogether. **(B)** Illumina-based sequencing was performed on each sample targeting V9 rDNA region. All reads were quality checked and dereplicated. Taxonomy assigned was done by homology (*ggsearch* global alignment) employing V9 PR2 reference database (de Vargas et al., 2015). From these, a total of 63,371 diatom assigned ribotypes (represented by ~14.6 million reads) were selected for global diatom distribution and diversity analyses. Classical morphology-based identification methods using light microscopy (LM) was done on few selected samples to validate the molecular data.

Figure 2. Photic-zone V9-rDNA diatom dataset.

(A) V9 rDNA rarefaction curve. *Upper panel*, a sample-based rarefaction curve (Coleman), representing V9 rDNA richness for diatom. *Lower panel*, each curve illustrates the estimated number of V9 rDNA (Coleman) for each ocean province. The color code for the ocean provinces is given under the figure. Notice the scale difference in x axis in the upper and lower panel. **(B)** Preston lognormal distribution of the diatom ribotype abundance in the entire data set. The number of unique diatom ribotypes is plotted for logarithmically binned abundance intervals. The part of the curve on the left of Preston's veil line (dashed black vertical line) corresponds to ribotypes with less than one read in the sample, and thus not represented in the dataset. The theoretical richness inferred from Preston Veil was found to be 96,710 ribotypes, indicating 33,339 ribotypes missed during the sampling. **(C)** Percentage contribution of diatoms to the total (i) eukaryotic, (ii) protistan, and (iii) photosynthetic planktonic community. The red-dashed line represent the mean percentage contribution by diatom to each of the indicated planktonic community. Each station label is color coded based on the province it belonged to. The lower right panel shows whether the sample was drawn (filled box) for the indicated depth and size class or not. Abbreviations: NAO – North Atlantic Ocean; MS – Mediterranean Sea; RS – Red Sea; IO – Indian Ocean; SAO – South Atlantic Ocean; SO – Southern Ocean; SPO – South Pacific Ocean. See Figure 1 for the location of station.

Figure 3. Summarizing diatom metabarcoding dataset.

All ribotypes were clustered based on their taxonomic affiliation at the level of genus and were organized under 76 genera plus 5 unassigned groups (Unassigned, Polar centric_X, Radial centric_X,

Raphid pennate_X, Araphid pennate_X). The color code for a genus is as follows: dark blue – polar centric; light blue – radial centric; dark green – raphid pennate; light green – araphid pennate; black – unassigned diatoms. The benthic or freshwater diatom genera are marked with asterix (**). **(a)** Abundance expressed as numbers of rDNA reads, **(b)** richness expressed as number of unique rDNA sequence and **(c)** the corresponding number of V9 rDNA OTUs are shown for each indicated genera. **(d)** Percentage distribution of rDNA reads per size class. **(e)** Percentage distribution of rDNA reads per depth. **(f)** Boxplot showing the mean percentage sequence similarity to a reference sequence. **(g)** Occupancy (*n*) expressed as the number of stations in which the genus was observed. The color codes for the four size class, two depths and occupancy are given under the figure.

Figure 4. Novelty inferred from *Tara* Oceans metabarcoding dataset.

(A) Percentage of unassigned ribotypes in each station (*left panel*). Within each station, 31-81% of the ribotypes could not be assigned to known diatom genera. The highest proportion of unassigned ribotypes was seen in Station 45 (~82%) followed by stations in the Pacific Ocean (Stations 109,110,111,122,123,124) (~63-66%). The most abundant stations (Stations 67 and 85) contained ~33% of unassigned ribotypes. Percentage of unassigned diatom community per depth in each province (*right panel*). **(B)** Percentage of unassigned ribotypes per size class. Surface samples corresponding to 0.8-5µm appear to have the highest percentage of unassigned ribotypes, whereas size fraction 20-180 has the lowest. **(C)** Percentage of unassigned ribotypes per size class in each province.

Figure 5. Comparisons of diatom community compositions estimated from V9 rDNA counts and by light microscopy.

Community composition profiles obtained from light microscopy and ribotype relative abundance inferred from taxonomy-based clustering of assigned ribotypes from eleven selected stations are shown.

Figure 6. Local and regional genus distribution and diversity inferred from *Tara* Oceans dataset.

(A) Distribution of top 20 diatom genera in seven oceanic provinces. These genera accounted for 98.3% of the assigned reads in the entire dataset. *Upper panel*, the variation in diversity for each indicated genus inferred from Shannon Diversity Index across 46 stations. *Pseudo-nitzchia*, *Chaetoceros* and *Thalassiosira* were the most diverse genera whereas *Corethron* and *Minidiscus* were among the least diverse. *Lower panel*, segments composing each bar are the percentage of reads in an ocean province for each indicated genus. *Chaetoceros*, *Corethron* and *Fragilariopsis* were abundant in the Southern Ocean. *Planktoniella* and *Haslea* were majorly seen in the Pacific Ocean. Genera are sorted by total

number of reads in the entire dataset. Bars are color coded by ocean province, as indicated. **(B)** Global distribution and diversity of the 10 most abundant genera. These genera accounted for 92.4% of the assigned reads in the entire dataset. The number of reads (*n*) assigned to each genus is indicated. The area of the bubble is scaled to the total number of reads for each genus at each location. For each panel, color key is shown in the legend. Red - low richness; Green - high richness.

Figure 7. Variation in diatom diversity across oceans.

(A) Variation in richness (expressed as unique number of ribotypes), effective number of species (ENS; expressed as exponentiated Shannon diversity index) and evenness across provinces. Number of stations sampled in each province are as follows: NAO – 1; MS – 12; RS – 4; IO – 11; SAO – 7; SO – 3; SPO – 9. **(B)** Diatom latitudinal diversity gradients. Shannon diversity index was computed for those stations (37 stations) for which surface samples for all size classes were available. Stations were grouped based on their latitudes in five major groups. The group median shows an increase of diversity towards lower latitudes. **(C)** Variation in diatom diversity across stations. Spatial variation of diatom diversity across 37 stations inferred from effective number of species (ENS; expressed as exponentiated Shannon diversity index). Low Shannon diversity indices were found in Stations 11, 45 and 84. The highest values of ribotype diversity were seen in Stations 65, 38 and 122. Each station (filled circle) is color coded based on the province it belonged to.

Figure 8. Cosmopolitanism, total abundance and station evenness of each diatom ribotype.

(A) Each bubble represent a ribotype (V9 rDNA); the radius being scaled to the number of reads it contains. The X-axis corresponds to the number of stations in which a ribotype occurs; the Y-axis corresponds to the evenness of the ribotype in those stations in which it occurs. The 20 most abundant ribotypes are labelled with their rank and assigned taxonomy. Many ribotypes showed high abundance (larger bubbles), low occupancy (x-axis) and low evenness (y-axis) For instance, ribotypes assigned to *Leptocylindrus* and *Corethron* (filled bubbles). Cosmopolitan ribotypes can be identified as those with highest occupancy. A range of evenness was exhibited by them. For instance, among the most abundant, ribotypes assigned to *Fragilariopsis*, *Chaetoceros* and *Thalassiosira* (filled bubbles) are cosmopolitan but with low evenness. This low evenness indicates that although present at all the stations, these ribotypes are dominant only in one or two. Ribotypes that could not be assigned to a genus level are indicated in red, to indicate the extent of undetermined diatom ribotypes. **(B)** Shared number of ribotypes among oceans. Bar graph showing the overlap cardinalities; sorted by overlap cardinality, presented from left to right, from greatest to least number of shared ribotypes. Counts are based on presence-absence. The color-coded numbers above the bars indicate the ribotypes exclusive to each province.

Figure 9. Biogeographic patterns.

Percentage of ribotypes shared between stations. Only those stations (37 stations) for which surface samples for all size classes were available are reported. For each station, a pooled community over size classes were obtained. Dendrogram of complete linkage clustering is shown. Pearson correlation was used as a distance measure to cluster stations. The five major clusters can be identified that represent five distinct latitudinal bands. Two major groups were identified, one with majority of stations from Atlantic, Pacific and Southern Oceans and another with Mediterranean Sea and low abundant stations from all oceanic regions. A substantial degree of sharing was seen among stations from Southern, Pacific and Mediterranean waters.

Supplementary Material

Supplementary Figure S1. Assessing V9 hypervariable sub-sequence (130 bp) of small-subunit (SSU) ribosomal DNA (rDNA) genes as diversity marker.

(A) 2,947 full length 18S rDNA diatom sequences obtained from PR2 reference database were used. Sequence variations along the entire 18S rDNA sequence were used to define nine hypervariable regions (V1-V9). Regions in red are V1-V9. The bases are numbered according to the alignment position. **(B-C)** Hypervariable region performance against the 18S rDNA sequence. Pairwise distances were calculated for 2497 diatom 18S rDNA sequences using Kimura-2-parameter model. Length variation and genetic distances for V1-V9 are shown. Regression of V1-V9 p-distance by NJ on to that of 18S sequence shows that V5 could better explain the phylogeny, followed by V4. Although the mean genetic distances were better in V4 and V9, they may not explain the phylogeny well. V9 performance was less than that of 18S. Taxa assignment at less than 70 % identity in V9 region is not recommended for diatoms. **(D)** Phylogenetic inference on Bacillariophyta from full-length 18S rDNA sequence phylogeny, V4 rDNA phylogeny, and V9 rDNA based phylogeny. Four prominent phylogenetic clades of diatoms, i.e. radial centric, polar centric, araphid pennate, raphid pennate are known. V4 and V9 sequences were used to check their performance in differentiating these four groups. Each of the hypervariable regions and full-length 18S rDNA sequences were aligned using MUSCLE and phylogenetic inference was done with NJ algorithm using pairwise distances in MEGA5. The tree was statistically tested using 1000 bootstrap.

Supplementary Figure S2. Novelty in *Tara* Oceans diatom data set.

Barplot showing the number of reference sequences present for each genus and the total number of unique V9 tags from *Tara* Oceans dataset assigned to it. The reference database has a total of 1,648 V9 sequences annotated as being derived from diatoms. The level of percentage identity to the reference sequence varied across ribotypes, but for this analysis a similarity cut-off of 85% was used. From a total of 63,371 ribotypes, 30,041 ribotypes were unassigned due to the lack of reference sequence.

Supplementary Figure S3. Photic-zone diatom OTU dataset.

(A) OTU rarefaction curve. A sample-based rarefaction curve (Coleman), representing OTU richness for diatom. **(B)** Estimating the completeness of sampling based on OTUs. OTU abundances were $\log_2$-transformed. Most of them were seen with intermediate abundances with a relatively few rare or very few ubiquitous OTUs. The area under the Preston curve provides an extrapolated estimate of richness and thus an indication of the completeness in the sampling effort. The theoretical OTU richness

inferred from Preston Veil was found to be 5,451 ribotypes, indicating 1,035 OTUs undetected.

Supplementary Figure S4. Comparing diatom distributions obtained from our study to the distribution reported in MAREDAT dataset.

(A) Venn diagram showing the overlap between *Tara* Oceans dataset, V9 PR2 reference database and MAREDAT. Green circle represents the subset of reference genera identified in the *Tara* Oceans dataset. **(B**) Coverage of MAREDAT database. **(C)** Coverage of *Tara* Oceans dataset**.**

Supplementary Figure S5. NMDS ordination of community obtained from "V9" and "LM" approach.

"V9" (red) and "LM" (blue) represent the results obtained from the genetic and microscopic methods, respectively.

Supplementary Figure S6. Global distribution and diversity of the genera ranked 11 to 20 based on their abundance.

The area of the bubble is scaled to the total number of reads for each genus at each location. For each panel, color key is shown in the legend. Red - low richness; Green - high richness.

Supplementary Figure S7. Variation in diversity per depth and size class.

(A) Ribotype richness. **(B)** Effective number of species (expressed as exponentiated SDI). **(C)** Evenness. The results indicate that 20-180 µm size fraction was the most diverse, showing higher diversity at DCM. The smaller 0.8-5 µm size fraction also showed a similar trend with higher level of diversity at DCM. The largest size fraction exhibited the lowest abundance and richness but have the highest evenness among all size fractions. In general, Surface was found to be less diverse and even than DCM samples.

Supplementary Figure S8. Variation in diatom diversity across stations.

Spatial variation of diatom diversity across 37 stations inferred from SDI and richness.

Supplementary Figure S9. Diatom ribotype composition based on incidence-based measure.

(A) Pairwise community dissimilarity (Bray-Curtis) across provinces, signifying higher dissimilarities for higher values. **(B)** Diatom ribotype composition (presence-absence). Pairwise Jaccard dissimilarity was used to cluster stations hierarchically (group-average linkage). A two-dimensional NMDS ordination is shown with a stress value of 0.19. Each station (filled circle) is color coded based on the province it belonged to.

Supplementary Figure S10. Diatom ribotype composition based on abundance-based measure.

(A) Pairwise Bray–Curtis distance was used to cluster stations hierarchically (group-average linkage). **(B)** The two-dimensional NMDS ordination of the transformed data in reduced space with a stress value of 0.16 was used to visualize pairwise Bray-Curtis distance among stations. Hellinger transformation was performed on the abundance matrix to minimize the influence of rare ribotypes. Each symbol corresponds to a station, colored based on provinces.

References

Allen AE, et al. (2011) Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. Nature 473:203-207.

Amaral-Zettler L, McCliment E, Huse S (2009) A method for studying protistan diversity using massively parallel sequencing of V9 hyper variable regions of small-subunit ribosomal RNA genes. PLoS ONE 4:e6372.

Amato A, Orsini L, D'Alelio D, Montresor M (2005) Lifecycle, size reduction patterns, and ultrastructure of the pennate planktonic diatom *Pseudo-nitzschia delicatissima* (Bacillariophyceae). J Phycol 41:542-556.

Armbrust EV (2009) The life of diatoms in the world's oceans. Nature 459:185-192.

Baas Becking LGM (1934) Geobiologie of inleiding tot de milieukunde. The Hague, the Netherlands: W.P. Van Stockum & Zoon (in Dutch).

Behnke A, Friedl T, Chepurnov VA, Mann DG (2004) Reproductive compatibility and rDNA sequence analyses in the *Sellaphora pupula* species complex (Bacillariophyceae). J Phycol 40:193-208.

Bellemain E, et al. (2010) ITS as an environmental DNA barcode for fungi: an *in silico* approach reveals potential PCR biases. BMC Microbiol 10:189.

Beszteri B, John U, Medlin LK (2007) An assessment of cryptic genetic diversity within the Cyclotella meneghiniana species complex (Bacillariophyta) based on nuclear and plastid genes, and amplified fragment length polymorphisms. An Eur J Phyco 42:47–60.

Bik HM, Halanych KM, Sharma J, Thomas WK (2012) Dramatic shifts in benthic microbial eukaryote communities following the deepwater horizon oil spill. PLoS ONE 7(6):e38550.

Bittner L, et al (2013) Diversity patterns of uncultured Haptophytes unravelled by pyrosequencing in Naples Bay. Mol Ecol 22:87–101.

Bopp L, Aumont O, Cadule P, Alvain S, Gehlen M (2005) Response of diatoms distribution to global warming and potential implications: A global model study. Geophys Res Lett 32:1–4.

Cermeño P, Falkowski PG (2009) Controls on diatom biogeography in the ocean. Science 325:1539–1541.

Cervato C, Burckle L (2003) Pattern of first and last appearance in diatoms: Oceanic circulation and the position of polar fronts during the Cenozoic. Paleoceanography 18:1055.

Chamnansinp A, Li Y, Lundholm N, Moestrup Ø (2013) Global diversity of two widespread, colony-forming diatoms of the marine plankton, *Chaetoceros socialis* (syn. *C. radians*) and *Chaetoceros gelidus sp. nov*. J Phycol 49:1128–1141.

Coissac E, Riaz T, Puillandre N (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. Mol Ecol 21(8):1834–1847.

Crosta X, Romero O, Armand LK, Pichon JJ (2005) The biogeography of major diatom taxa in Southern Ocean sediments: 2. Open ocean related species. Palaeogeogr Palaeoclimatol Palaeoecol 223:66-92.

Cunningham SA, Alderson SG, King BA (2003) Transport and variability of the Antarctic Circumpolar Current in Drake Passage. J Geophys Res 108:8084.

Dayton PK, Morbida BJ, Bacon F (1994) Polar marine communities. Amer Zool:3490-3499.

de Vargas C, et al. (2015) Eukaryotic plankton diversity in the sunlit global ocean. Science (In press).

Degerlund M, Huseby S, Zingone A, Sarno D, Landfald B (2012) Functional diversity in cryptic species of *Chaetoceros socialis Lauder* (Bacillariophyceae). J Plankton Res 34:416–431.

Evans KM, Hayes PK (2004) Microsatellite markers for the cosmopolitan marine diatom *Pseudo-nitzschia pungens*. Mol Ecol Notes 4:125–126.

Falkowski PG (2002) The ocean's invisible forest - marine phytoplankton play a critical role in regulating the earth's climate. Could they also be used to combat global warming. Sci Am 287(2):54–61.

Falkowski PG, Barber RT, Smetacek V (1998) Biogeochemical controls and feedbacks on ocean primary production. Science 281(5374):200–206.

Fernandes LF, Calixto-Feres M (2012) Morphology and distribution of two epizoic diatoms (Bacillariophyta) in Brazil. Acta Bot Bras 26(4):836-841.

Ficetola GF, et al (2010) An *in silico* approach for the evaluation of DNA barcodes. BMC Genom 11:434.

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P (1998) Primary production of the biosphere: Integrating terrestrial and oceanic components. Science 281:237-240.

Fourtanier E, Kociolek JP (2003) Catalogue of the diatom genera (vol 14, pg 190, 1999). Diatom Res 18:245–258.

Fuhrman JA, et al. (2008) A latitudinal diversity gradient in planktonic marine bacteria. Proc Natl Acad Sci USA 105(22):7774-7778.

Gallagher JC (1980) Population genetics of *Skeletonema costatum* (Bacillariophyceae) in Narragansett bay. J Phycol 16:464-474.

Galluzzi L, et al. (2004) Development of a real-time PCR assay for rapid detection and quantification of *Alexandrium minutum* (a dinoflagellate). Appl Environ Microbiol 70:1199–1206.

Garate-Lizarraga I, Muneton-Gomez MDS (2009) Primer registro de la diatomea epibionte *Pseudohimantidium pacificum* y de otras asociaciones simbióticas en el Golfo de California. Act Bot Mex (88):31-45.

Gersonde R, Zielinski U (2000) The reconstruction of late quaternary antarctic sea-ice distribution - the use of diatoms as a proxy for sea-ice. Palaeogeogr Palaeoclimatol Palaeoecol 162(3-4):263-286.

Godhe A, et al. (2008) Quantifying diatom and dinoflagellate biomass in coastal marine sea water samples by real-time PCR. Appl Environ Microbiol 74:7174–7182.

Guillou L, et al. (2013) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Res 41(D1):D597-D604.

Guiry MD (2012) How many species of algae are three? J Phycol 48:1057–1063.

Hill MO (1973) Diversity and evenness: a unifying notation and its consequences. Ecology 54:427-432.

Hinder SL, et al. (2012) Changes in marine dinoflagellate and diatom abundance under climate change. Nature Clim Change 2:271–275.

Huseby S, Degerlund M, Zingone A, Hansen E (2012) Metabolic fingerprinting reveals differences between northern and southern strains of the cryptic diatom *Chaetoceros socialis*. Eur J Phycol 47:480–489.

Janech M , Krell A, Mock T , Kang JS and Raymond J (2006) Ice-bindingproteins from sea ice diatoms (Bacillariophyceae). J Phycol 42(2):410-416.

Ji Y, et al. (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecol Lett 16(10):1245–1257.

Karsenti E, et al. (2011) A holistic approach to marine eco-systems biology. PLoS Biol 9:e1001177.

Ki JS, Han MS (2005) Molecular analysis of complete SSU to LSU rDNA sequence in the harmful dinoflagellate *Alexandrium tamarense* (Korean isolate, HY970328M). Ocean Sci J 40:155-166.

Kröger N, Sumper M (2000) The Biochemistry of silica formation in diatoms. *Biomineralization*, ed Baeuerlein E (Wiley-VCH, Weinheim), pp 151-170.

Leblanc K, et al. (2012) A global diatom database – abundance, biovolume and biomass in the world ocean. Earth Syst Sci Data 4:149-165.

Lundholm N, et al. (2006) Inter- and intraspecific variation of the *Pseudo-nitzschia delicatissima*-complex (Bacillariophyceae) illustrated by rRNA probes, morphological data and phylogenetic analyses. J Phycol 42:464-481.

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.

Mann DG, Droop SJM (1996) Biodiversity, biogeography and conservation of diatoms. Hydrobiologia 336:19–32.

Mann DG, Vanormelingen P (2013) An inordinate fondness? The number, distributions, and origins of diatom species. J Euk Microbiol 60(4):414-420.

Meusnier I, et al. (2008) A universal DNA mini-barcode for biodiversity analysis. BMC Genomics 9(1):214.

Nanjappa D, Audic S, Romac S, Kooistra WH, Zingone A (2014). Assessment of species diversity and distribution of an ancient diatom lineage using a DNA metabarcoding approach. PloS ONE 9(8):e103810.

Nelson DM, Treguer P, Brzezinski MA, Leynaert A, Queguiner B (1995) Production and dissolution of biogenic silica in the ocean - revised global estimates, comparison with regional data and relationship to biogenic sedimentation. Global Biogeochem Cy 9:359-372.

Nolte V, et al. (2010) Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. Mol Ecol 19:2908–2915.

Pesant et al., (2015) Open science resources for the discovery and analysis of the Tara Oceans Data Collection. Sci Data (In press).

Peterson RJ, Stramma L (1991) Upper level circulation in the South Atlantic Ocean. Prog Oceanogr 26(1):1–73.

Pielou E (1966) The measurement of diversity in different types of biological collections. J Theor Biol 13:131-144.

R Development Core Team (2009) R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Riaz T, et al. (2011) ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. Nucleic Acids Res 39(21):e145.

Ribera d'Alcala M, et al. (2004) Seasonal patterns in plankton communities in a pluriannual time series at a coastal Mediterranean site (Gulf of Naples): an attempt to discern recurrences and trends. Sci Mar 68 (Suppl. 1):65–83.

Rodríguez-Martínez R, Gabrielle R, Guillem S, Ramon M (2013) Biogeography of the uncultured marine picoeukaryote MAST-4: temperature-driven distribution patterns. ISMEJ 7:531–1543.

Round FE, Crawford RM, Mann DG (1990) The Diatoms. Biology and Morphology of the Genera. Cambridge University Press, Cambridge, 747 pp.

Rynearson TA, Armbrust EV (2000) DNA fingerprinting reveals extensive genetic diversity in a field population of the centric diatom *Ditylum brightwellii*. Limnol Oceanogr 45:1329-1340.

Sarno D, Kooistra WCHF, Medlin LK, 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. J Phycol 41:151-176.

Siedler G, Griffies S, Gould J, Church J (2013) Ocean Circulation and Climate: A 21st century perspective. Academic Press 904 pp.

Siokou-Frangou I, et al. (2010) Plankton in the open Mediterranean Sea: a review. Biogeosciences 7(5):1543-1586.

Skov J, Lundholm N, Pocklington R, Rosendahl S, Moestrup O (1997) Studies on the marine planktonic diatom *Pseudo-nitzschia*. 1. Isozyme variation among isolates of *P. pseudodelicatissima* during a bloom in Danish coastal waters. Phycologia 36:374–380.

Skovgaard A, Saiz E (2006) Seasonal occurrence and role of protistan parasites in coastal marine zooplankton. Marine Ecology - Progress Series, vol 327, pp. 37-49.

Slamovits CH, Okamoto N, Burri L, James ER, Keeling PJ (2011) A bacterial proteorhodopsin proton pump in marine eukaryotes. Nat Commun 2:1–6.

Smetacek V (1998) Diatoms and the silicate factor. Nature 391:224-225.

Smol JP, Stoermer EF (eds.) (2010) The Diatoms: Applications for Environmental and Earth Sciences (Cambridge University Press), pp 667.

Sogin ML, et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci USA 103(32):12115–12120.

Soudek DJR, Robinson GGC (1983) Electrophoreticanalysis of the species and population structure of the diatom *Asterionella formosa*. Can J Bot 61:418–433.

Sournia A, Chrdtiennot-Dinet MJ, Ricard M (1991) Marine phytoplankton: how many species in the world ocean? J Plankton Res 13(5):1093-1099.

Steele PR, Pires JC (2011) Biodiversity assessment: state-of-the-art techniques in phylogenomics and species identification. Am J Bot 98(3):415–425.

Sul WJ, Oliver TA, Ducklow HW, Amaral-Zettler LA, Sogin ML (2013) Marine bacteria exhibit a bipolar distribution. Proc Natl Acad Sci USA 110(6):2342-7.

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. Mol Ecol 21:2045-2050.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol Biol Evol 30:2725-2729.

Vanormelingen P, Verleyen E, Vyverman W (2008) The diversity and distribution of diatoms: from cosmopolitanism to narrow endemism. Biodiversity Conserv 17:393–405.

Villar E, et al. (2015) Environmental disturbance in Agulhas rings affect inter-ocean plankton dispersal. Science. (In press)

Will KW, Mishler BD, Wheeler QD (2005) The perils of DNA barcoding and the need for integrative taxonomy. Syst Biol 54(5):844–851.

Yoccoz NG (2012) The future of environmental DNA in ecology. Mol Ecol 21:2031-2038.

Yu DW, et al. (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods Ecol Evol 3:613–623.

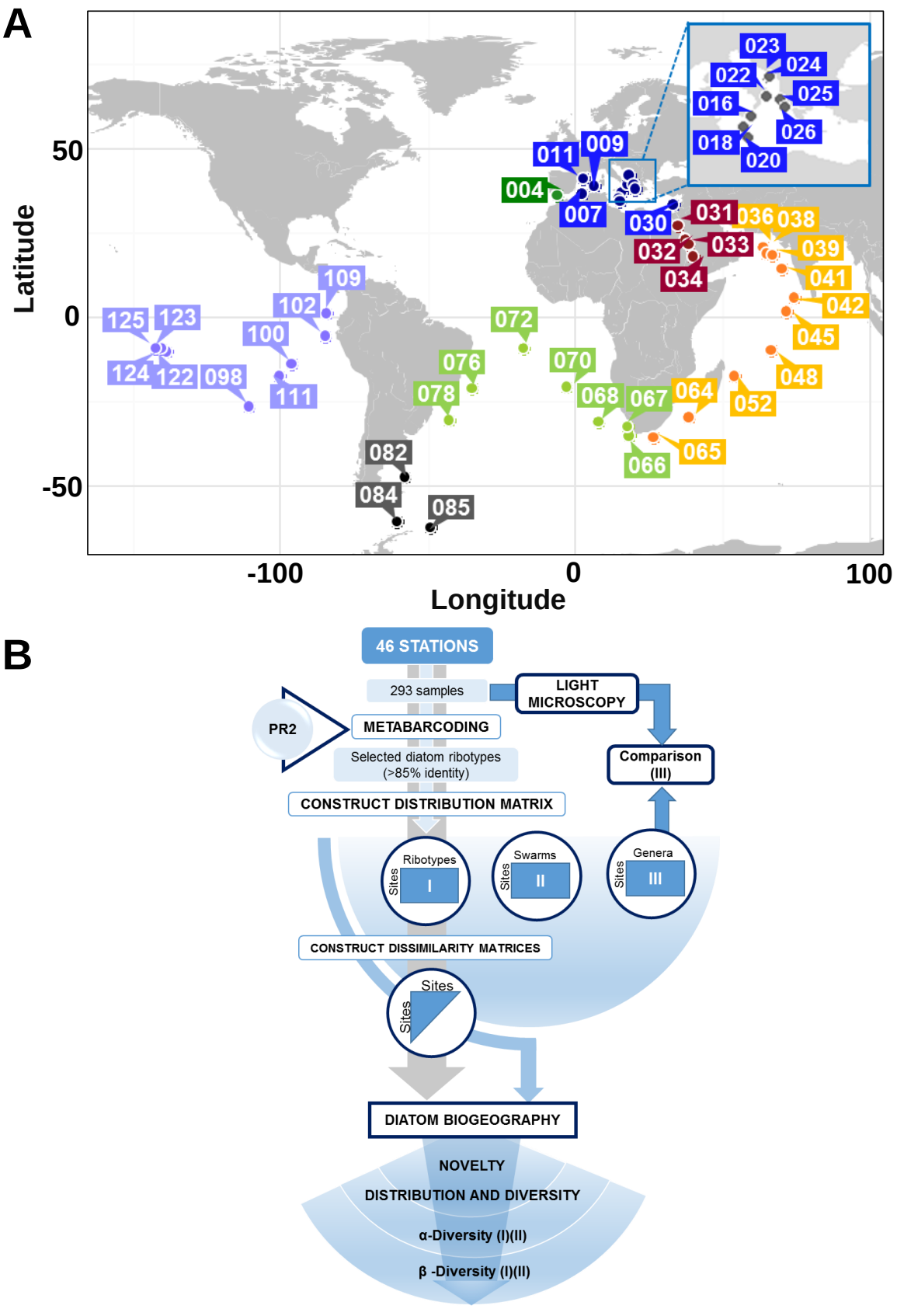

A
Latitude
50
0
-50
-100
0
100
Longitude
023
024
022
025
016
026
018
020
011
009
004
007
030
031
036
038
032
033
039
034
041
042
045
109
102
125
123
100
072
076
070
124
122
098
078
068
067
064
048
052
111
065
082
066
084
085
B
46 STATIONS
293 samples
LIGHT MICROSCOPY
PR2
METABARCODING
Selected diatom ribotypes (>85% identity)
Comparison (III)
CONSTRUCT DISTRIBUTION MATRIX
Ribotypes
Sites
I
Swarms
Sites
II
Genera
Sites
III
CONSTRUCT DISSIMILARITY MATRICES
Sites
Sites
DIATOM BIOGEOGRAPHY
NOVELTY
DISTRIBUTION AND DIVERSITY
α-Diversity (I)(II)
β -Diversity (I)(II)

Figure 1

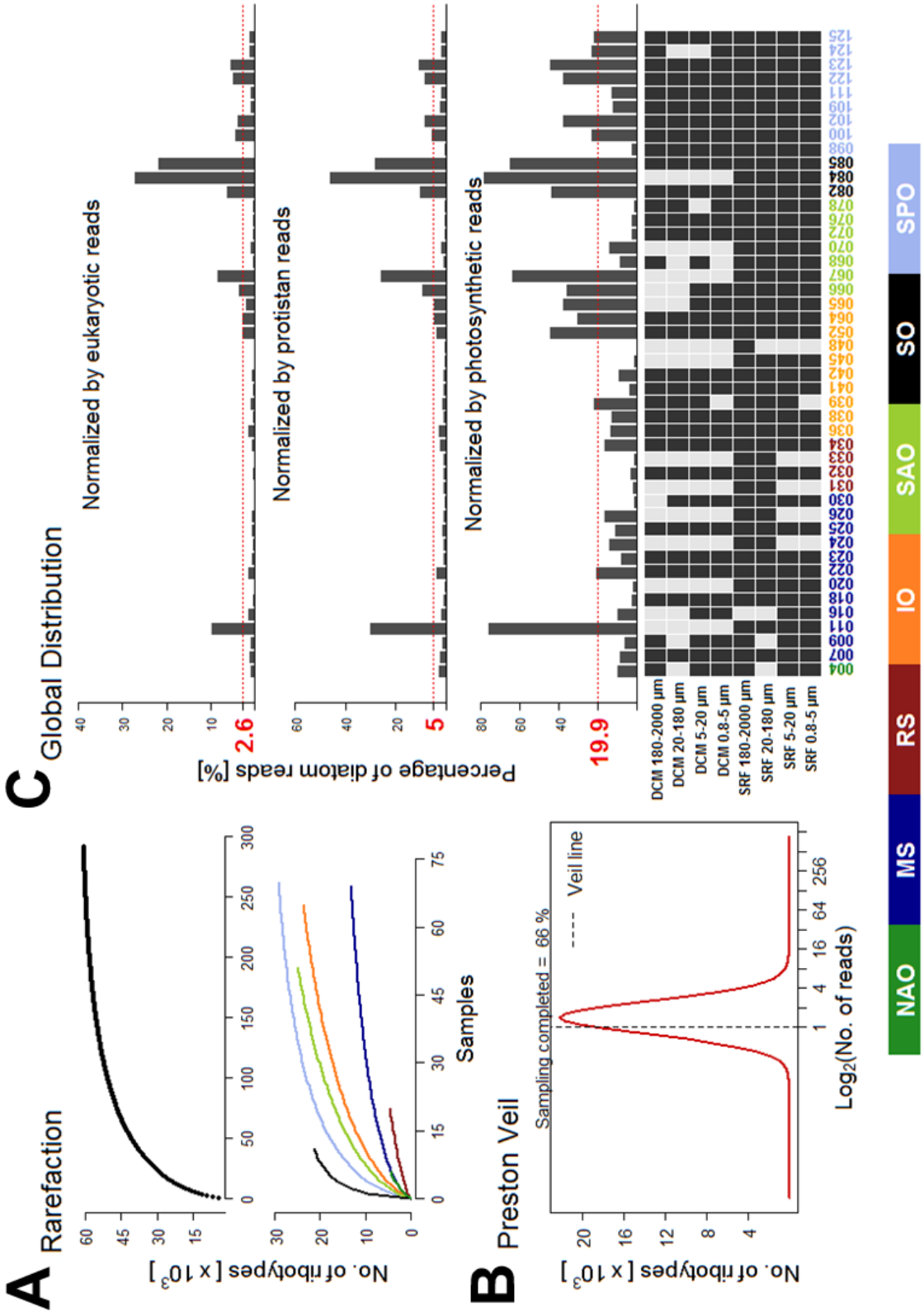

A Rarefaction
No. of rRNA types [x 10³]
Samples
B Preston Veil
Sampling completed = 66 %
Veil line
No. of rRNA types [x 10³]
Log2(No. of reads)
C Global Distribution
Normalized by eukaryotic reads
Normalized by protistan reads
Normalized by photosynthetic reads
Percentage of diatom reads [%]
2.6
5
19.9
DCM 180-2000 µm
DCM 20-180 µm
DCM 5-20 µm
DCM 0.8-5 µm
SRF 180-2000 µm
SRF 20-180 µm
SRF 5-20 µm
SRF 0.8-5 µm
NAO
MS
RS
IO
SAO
SO
SPO

Figure 2

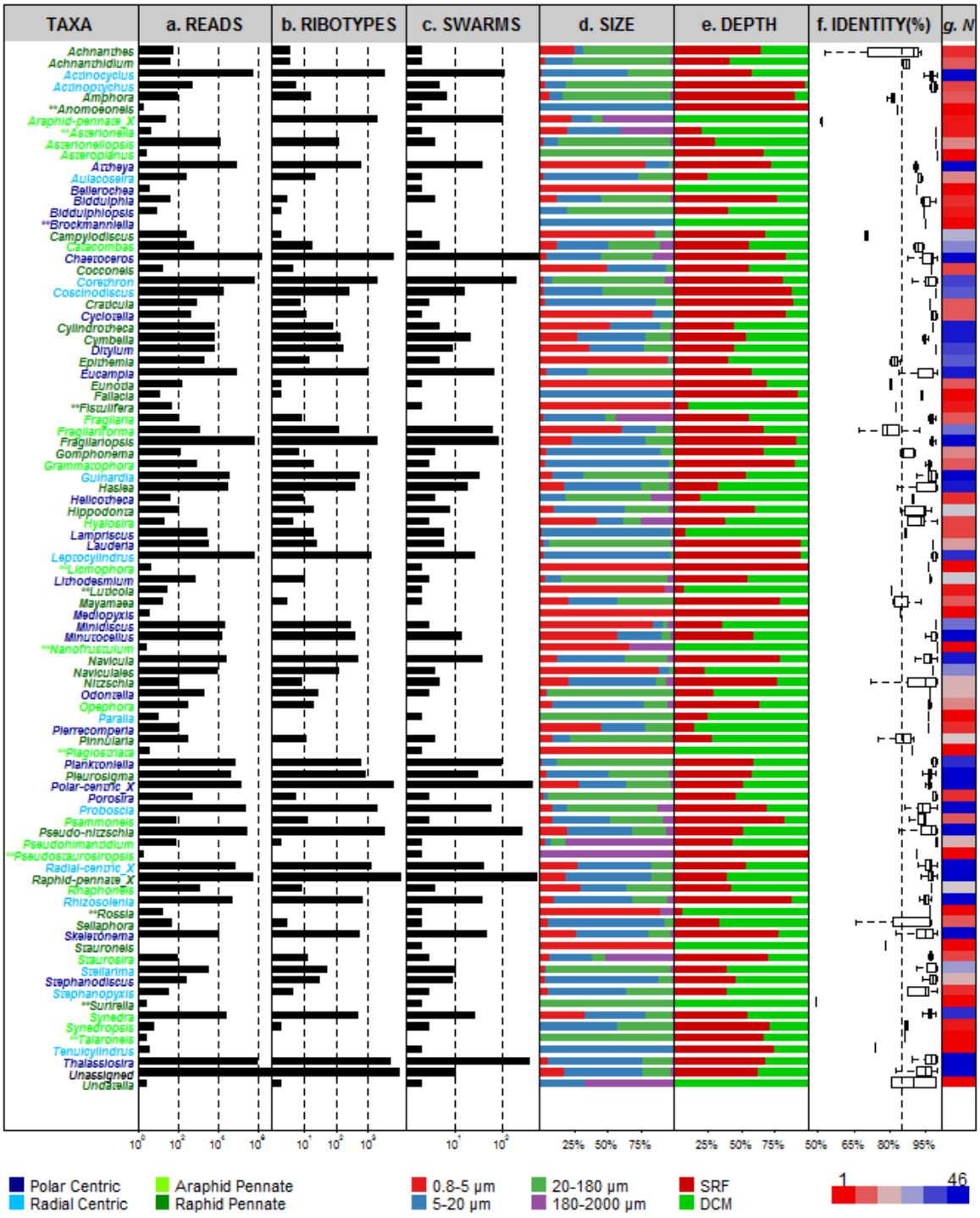

TAXA
a. READS
b. RIBOTYPES
c. SWARMS
d. SIZE
e. DEPTH
f. IDENTITY(%)
g. N
Achnanthes
Achnanthidium
Actinocyclus
Actinoptychus
Amphora
**Anomoeoneis
Araphid-pennate_X
**Asterionella
Asterionellopsis
Asteroplanus
Attheya
Aulacoseira
Bellerochea
Biddulphia
Biddulphiopsis
**Brockmanniella
Campylodiscus
Catacombas
Chaetoceros
Cocconeis
Corethron
Coscinodiscus
Cradleula
Cyclotella
Cylindrotheca
Cymbella
Ditylum
Epithemia
Eucampia
Eunotia
Fallacia
**Fistulifera
Fragilaria
Fragilariforma
Fragilariopsis
Gomphonema
Grammatophora
Guinardia
Haslea
Helicotheca
Hippodonta
Hyalosira
Lampriscus
Lauderia
Leptocylindrus
**Licmophora
Lithodesmium
**Luticola
Mayamaea
Mediopyxis
Minidiscus
Minutocellus
**Nanofrustulum
Navicula
Naviculales
Nitzschia
Odontella
Opephora
Paralia
Pierrecomperia
Pinnularia
**Plagiostriata
Planktoniella
Pleurosigma
Polar-centric_X
Porosira
Proboscia
Psammoneis
Pseudo-nitzschia
Pseudohimantidium
**Pseudostaurosiropsis
Radial-centric_X
Raphid-pennate_X
Rhaphoneis
Rhizosolenia
**Rossia
Sellaphora
Skeletonema
Stauroneis
Staurosira
Stellarima
Stephanodiscus
Stephanopyxis
**Sunirella
Synedra
Synedropsis
**Talaroneis
Tenuicylindrus
Thalassiosira
Unassigned
Undatella
Polar Centric
Radial Centric
Araphid Pennate
Raphid Pennate
0.8-5 µm
5-20 µm
20-180 µm
180-2000 µm
SRF
DCM
1
46

Figure 3

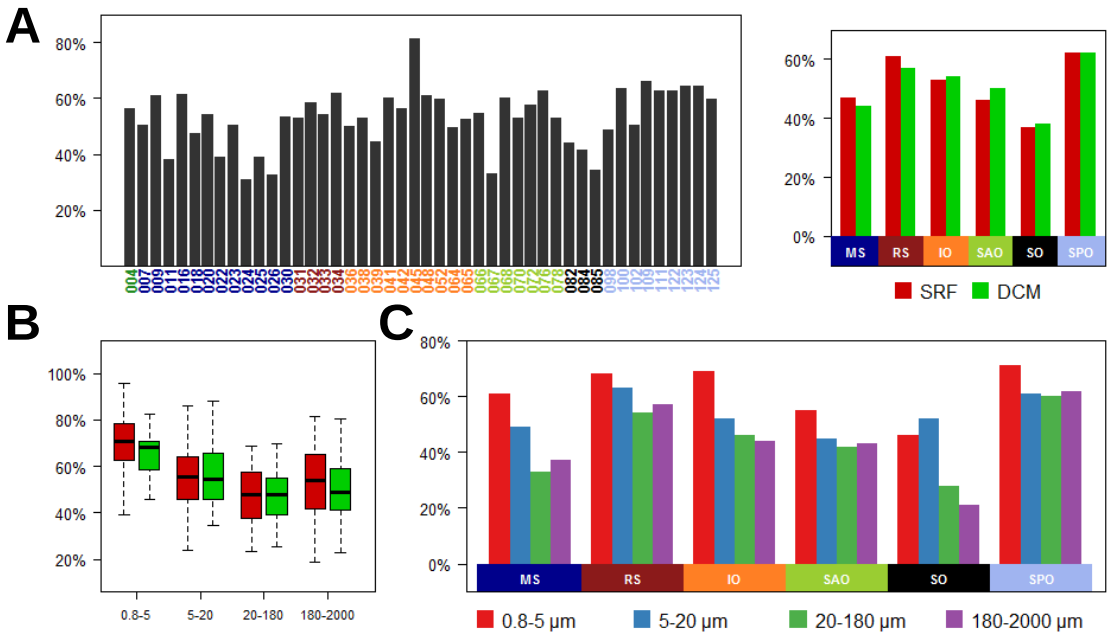

A
80%
60%
40%
20%
004 007 009 011 016 018 020 022 023 024 026 030 031 032 033 034 036 038 039 041 042 048 049 052 054 065 066 067 068 070 072 076 078 082 084 085 093 102 105 111 122 123 124 125
60%
40%
20%
0%
MS
RS
IO
SAO
SO
SPO
SRF
DCM
B
100%
80%
60%
40%
20%
0.8-5
5-20
20-180
180-2000
C
80%
60%
40%
20%
0%
MS
RS
IO
SAO
SO
SPO
0.8-5 µm
5-20 µm
20-180 µm
180-2000 µm

Figure 4

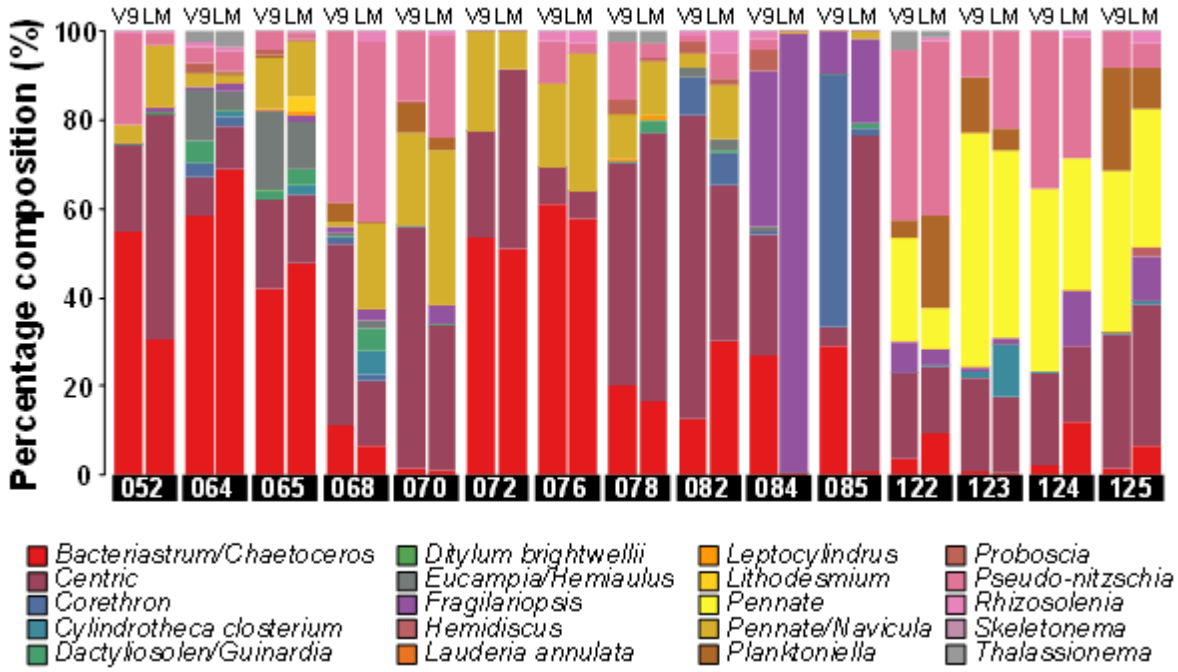

V9 LM
Percentage composition (%)
100
80
60
40
20
0
052
064
065
068
070
072
076
078
082
084
085
122
123
124
125
Bacteriastrum/Chaetoceros
Centric
Corethron
Cylindrotheca closterium
Dactyliosolen/Guinardia
Ditylum brightwellii
Eucampia/Hemiaulus
Fragilariopsis
Hemidiscus
Lauderia annulata
Leptocylindrus
Lithodesmium
Pennate
Pennate/Navicula
Planktoniella
Proboscia
Pseudo-nitzschia
Rhizosolenia
Skeletonema
Thalassionema

Figure 5

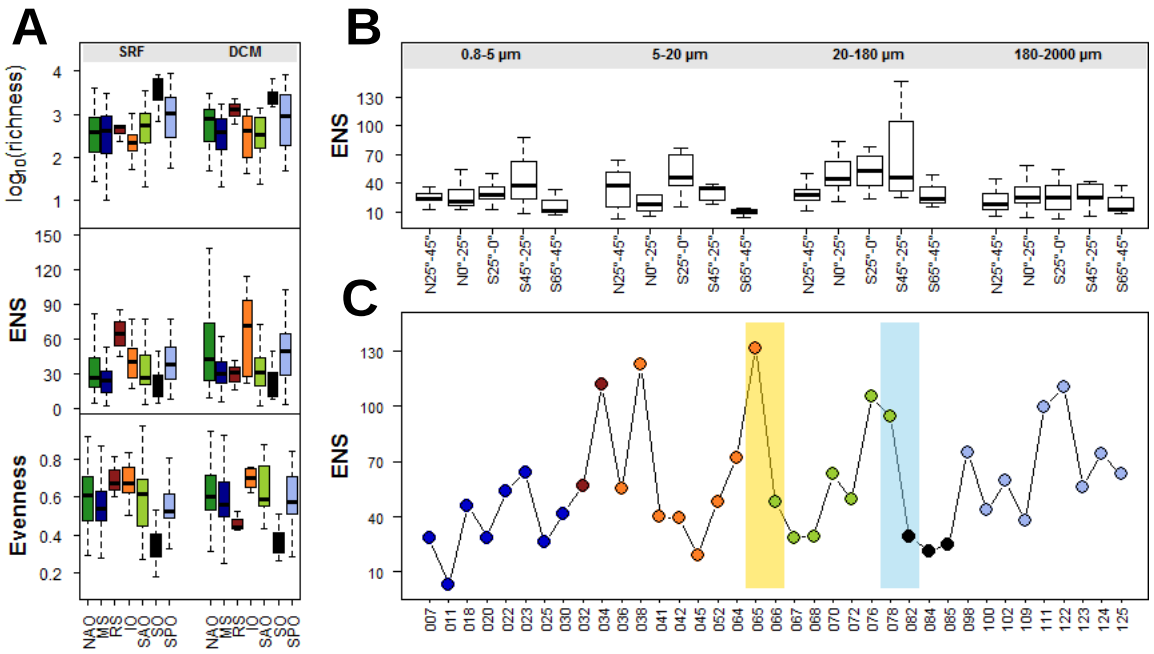

A
SRF
DCM
log10(richness)
ENS
Evenness
NAO
MS
RS
IO
SAO
SO
SPO
B
0.8-5 µm
5-20 µm
20-180 µm
180-2000 µm
ENS
N25°-45°
N0°-25°
S25°-0°
S45°-25°
S65°-45°
C
ENS
007
011
018
020
022
023
025
030
032
034
036
038
041
042
045
052
064
065
066
067
068
070
072
076
078
082
084
085
098
100
102
109
111
122
123
124
125

Figure 6

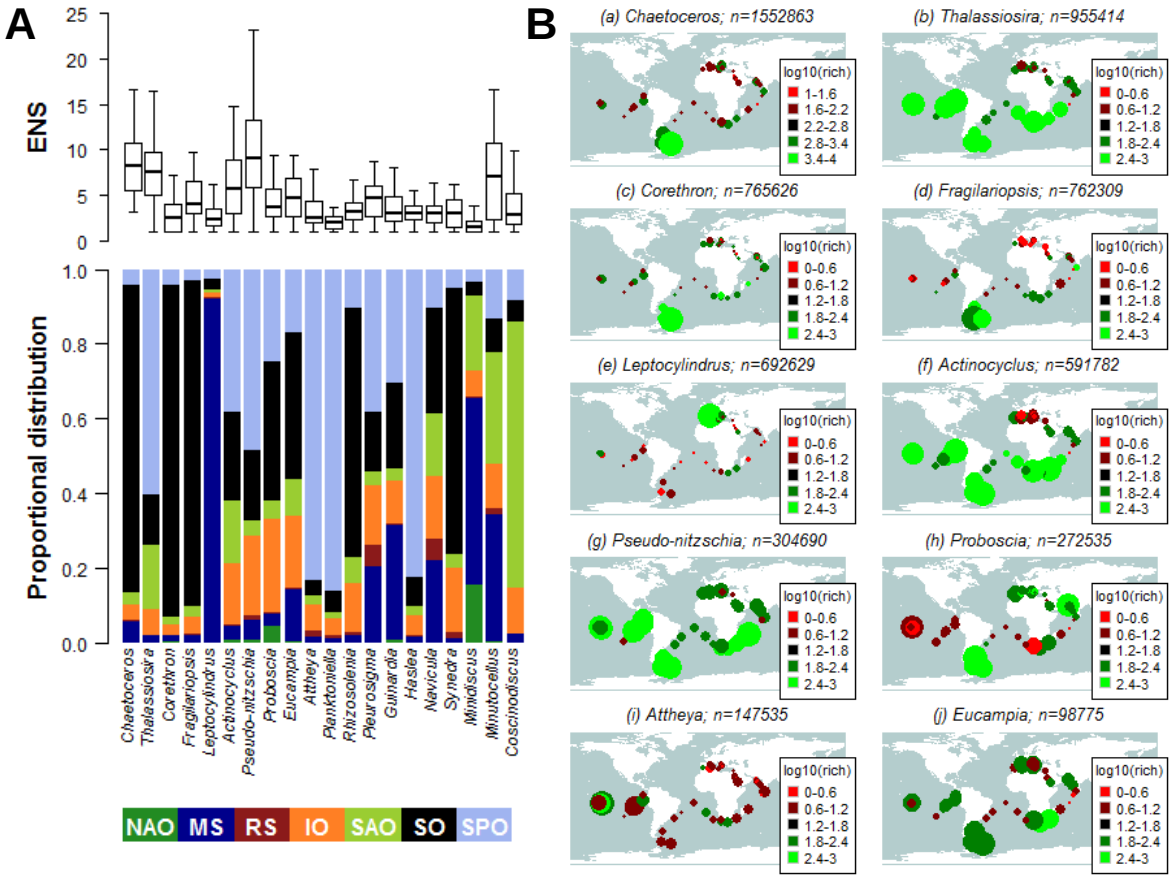

A
ENS
Proportional distribution
Chaetoceros
Thalassiosira
Corethron
Fragilariopsis
Leptocylindrus
Actinocyclus
Pseudo-nitzschia
Proboscia
Eucampia
Attheya
Planktoniella
Rhizosolenia
Pleurosigma
Guinardia
Hastea
Navicula
Synedra
Minidiscus
Minutocellus
Coscinodiscus
NAO
MS
RS
IO
SAO
SO
SPO
B
(a) Chaetoceros; n=1552863
(b) Thalassiosira; n=955414
(c) Corethron; n=765626
(d) Fragilariopsis; n=762309
(e) Leptocylindrus; n=692629
(f) Actinocyclus; n=591782
(g) Pseudo-nitzschia; n=304690
(h) Proboscia; n=272535
(i) Attheya; n=147535
(j) Eucampia; n=98775
log10(rich)

Figure 7

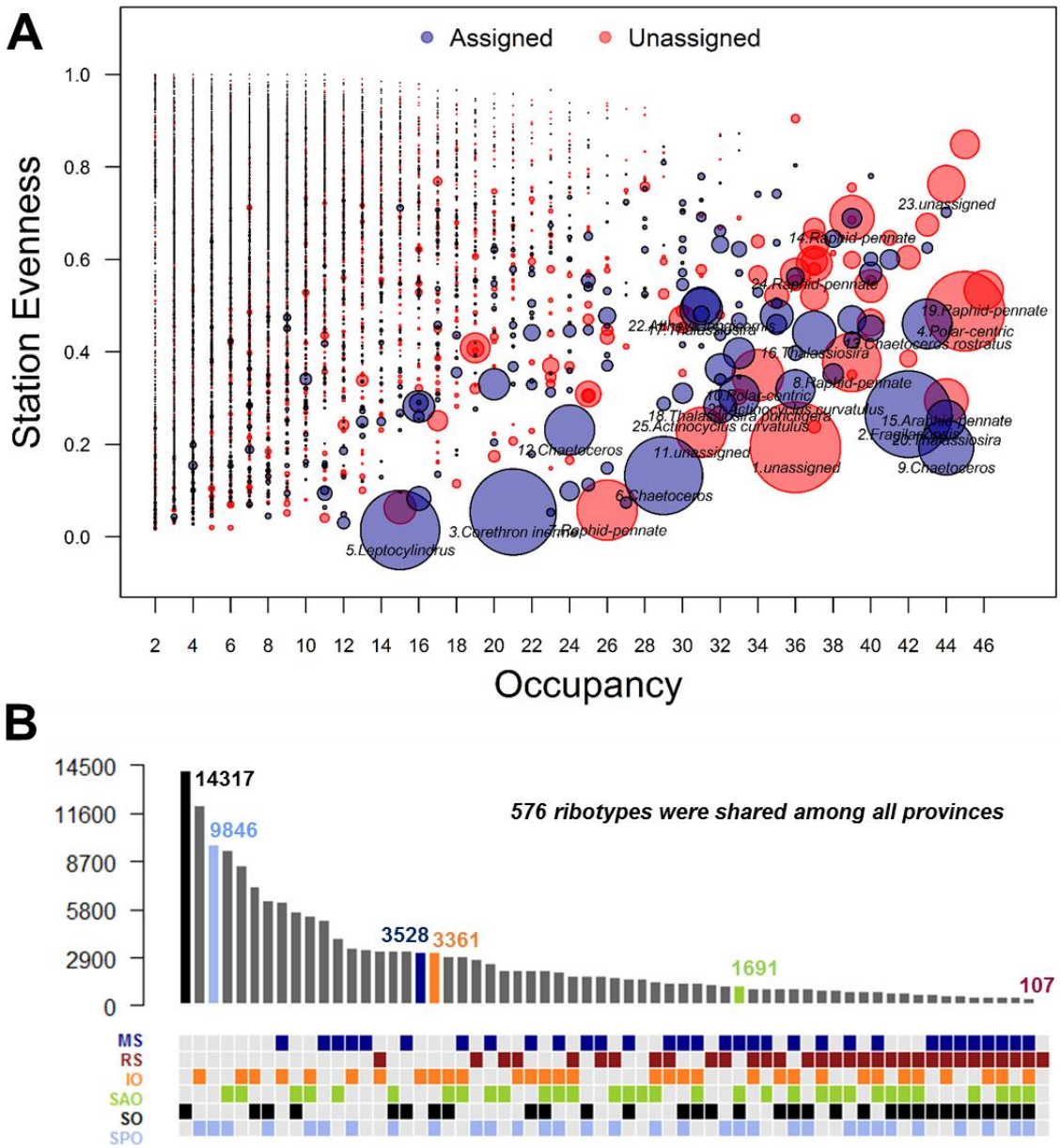

A
Assigned
Unassigned
Station Evenness
1.0
0.8
0.6
0.4
0.2
0.0
2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46
Occupancy
23.unassigned
14.Raphid-pennate
24.Raphid-pennate
19.Raphid-pennate
4.Polar-centric
16.Thalassiosira
8.Raphid-pennate
15.Araphid-pennate
9.Chaetoceros
1.unassigned
11.unassigned
12.Chaetoceros
6.Chaetoceros
3.Corethron inerme
Raphid-pennate
5.Leptocylindrus
25.Actinocyclus curvatulus
B
14500
11600
8700
5800
2900
0
14317
9846
3528
3361
1691
107
576 ribotypes were shared among all provinces
MS
RS
IO
SAO
SO
SPO

Figure 8

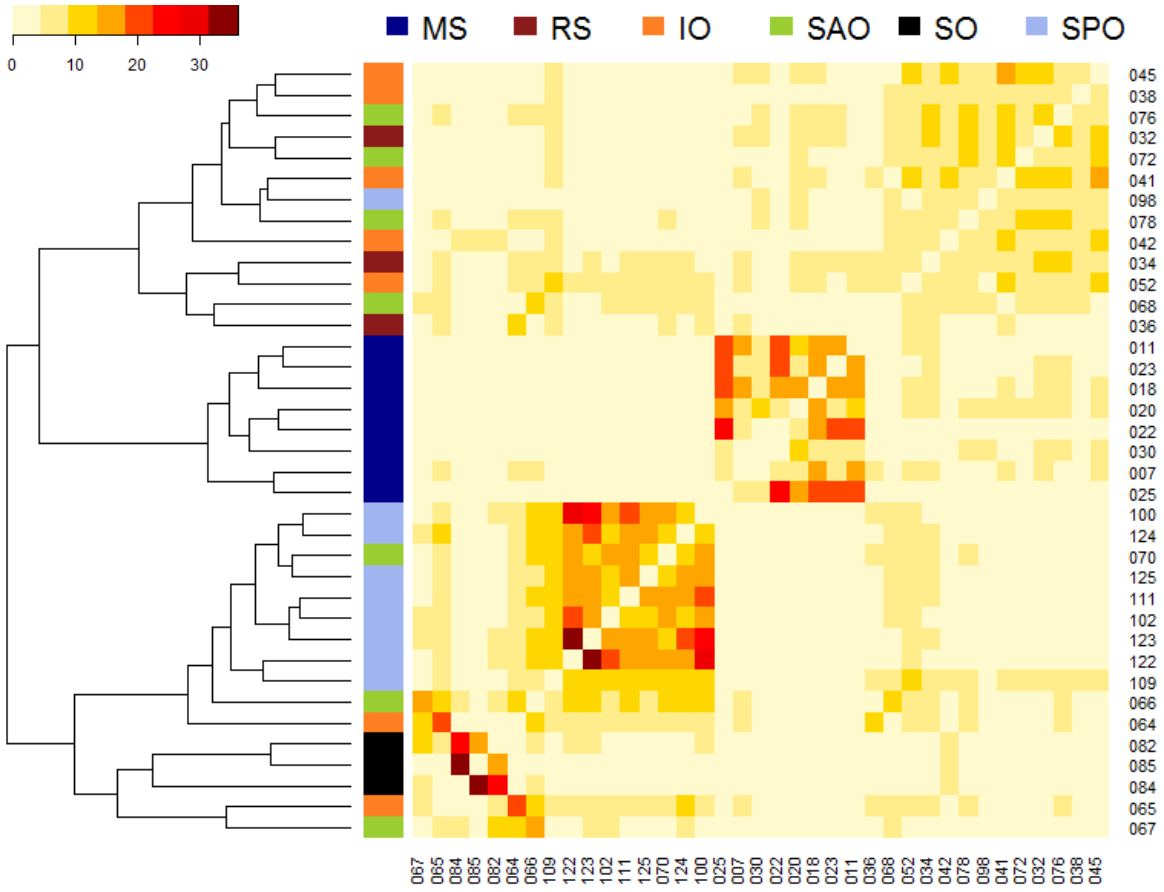
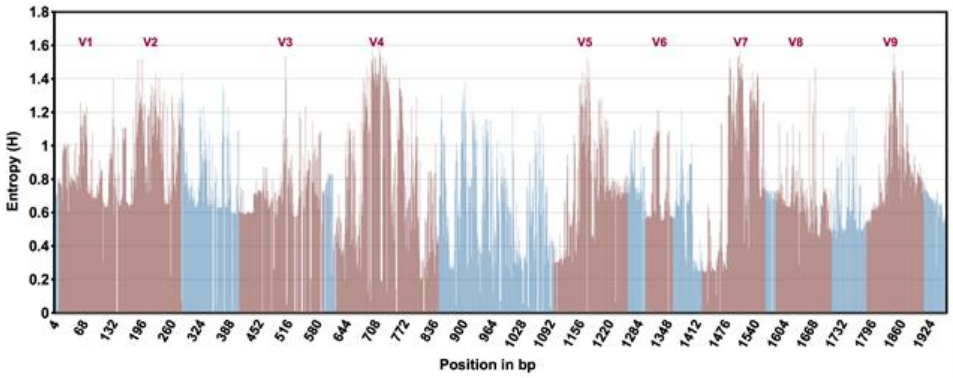

0
10
20
30
MS
RS
IO
SAO
SO
SPO
045
038
076
032
072
041
098
078
042
034
052
068
036
011
023
018
020
022
030
007
025
100
124
070
125
111
102
123
122
109
066
064
082
085
084
065
067
067
065
084
085
082
064
066
109
122
123
102
111
125
070
124
100
025
007
030
022
020
018
023
011
036
068
052
034
042
078
098
041
072
032
076
038
045

Figure 9

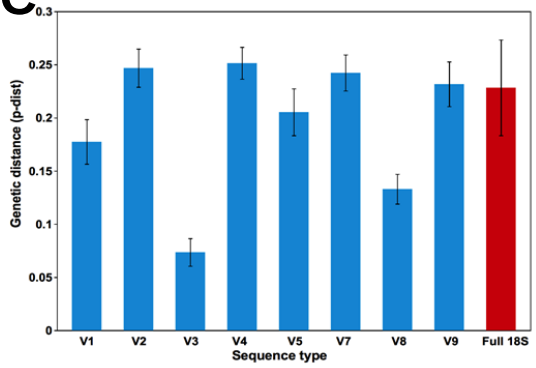
A



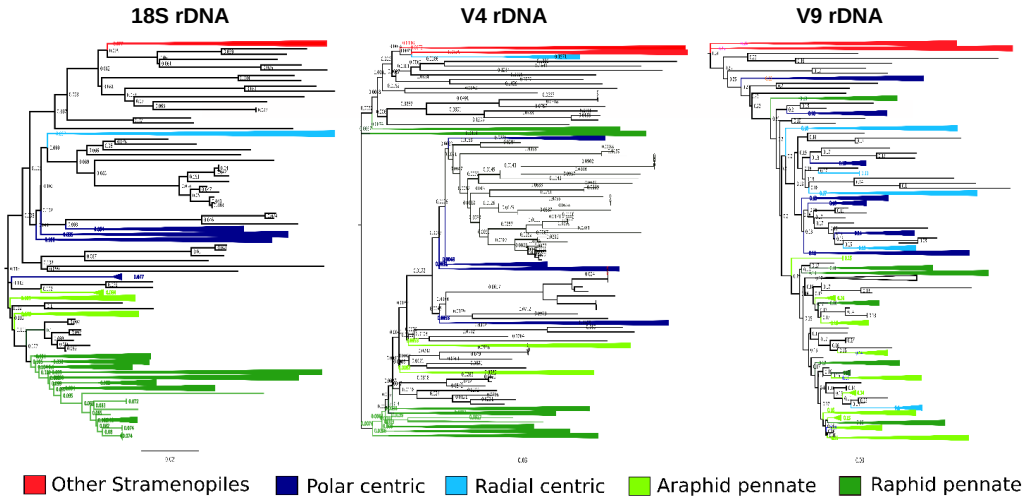
B

Hyper-variable region	Regression value (R^2) on all 18S diatom	Position in bp (*P. tricornutum* as reference)	Number of sequences	Minimum percentage of identity to differentiate two species
V1	0.087	3 - 103	1664	84%
V2	0.250	92 - 290	1853	86%
V3	0.071	401 - 561	2204	86%
V4	0.070	588 - 781	2799	74%
V5	0.137	1003 - 1154	2815	87%
V7	0.264	1302 - 1565	2756	80%
V8	0.256	1448 - 1565	2501	84%
V9	0.219	1640 - 1755	1777	70%
18S	-	1 - 1767	2947	93%

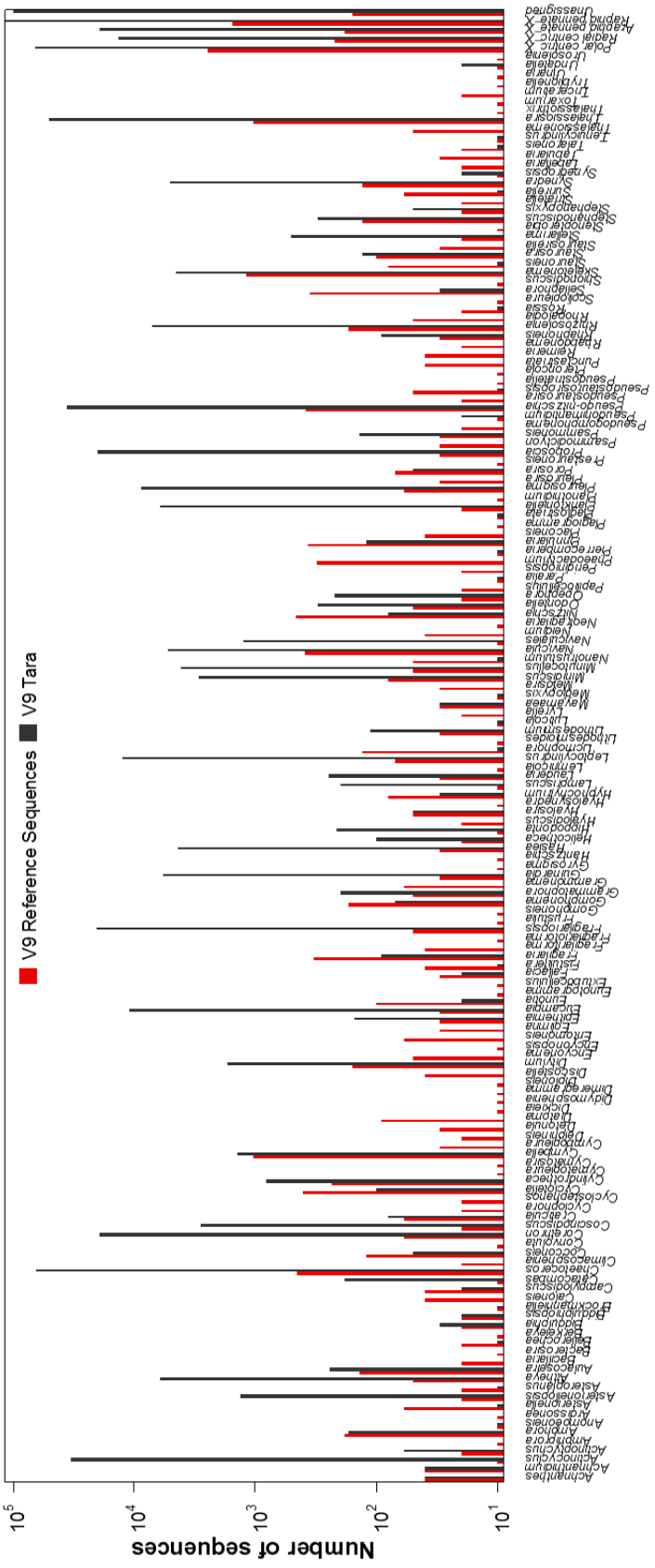
C



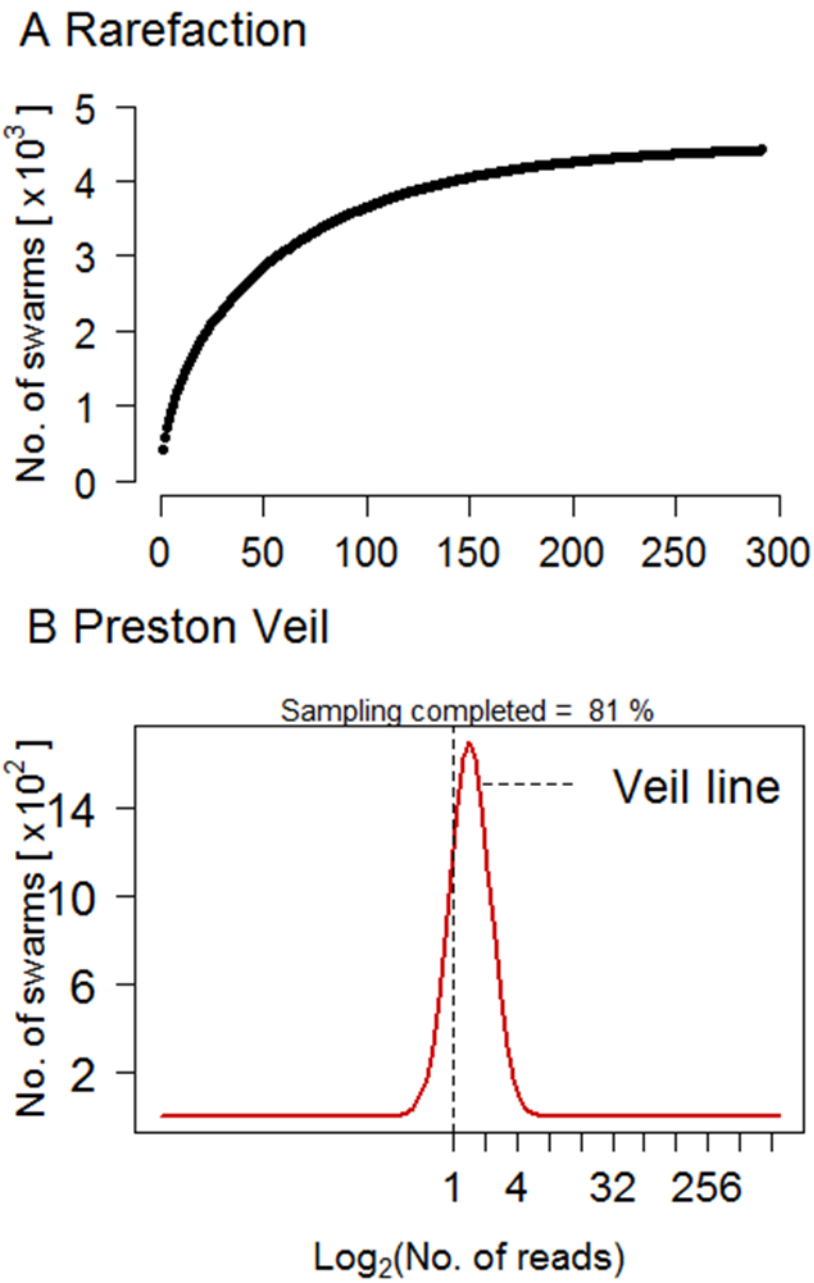
D



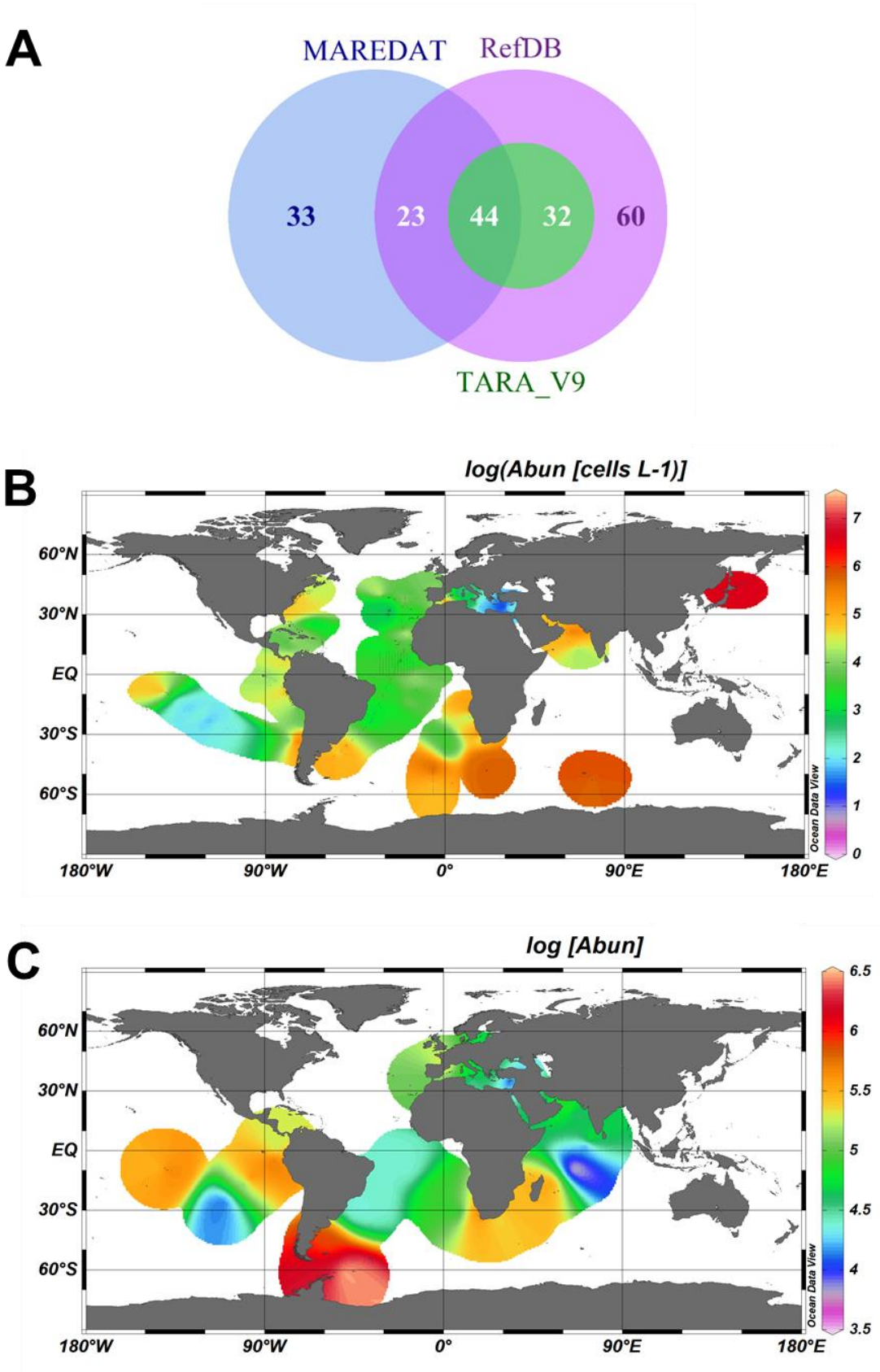
Supplementary figure 1



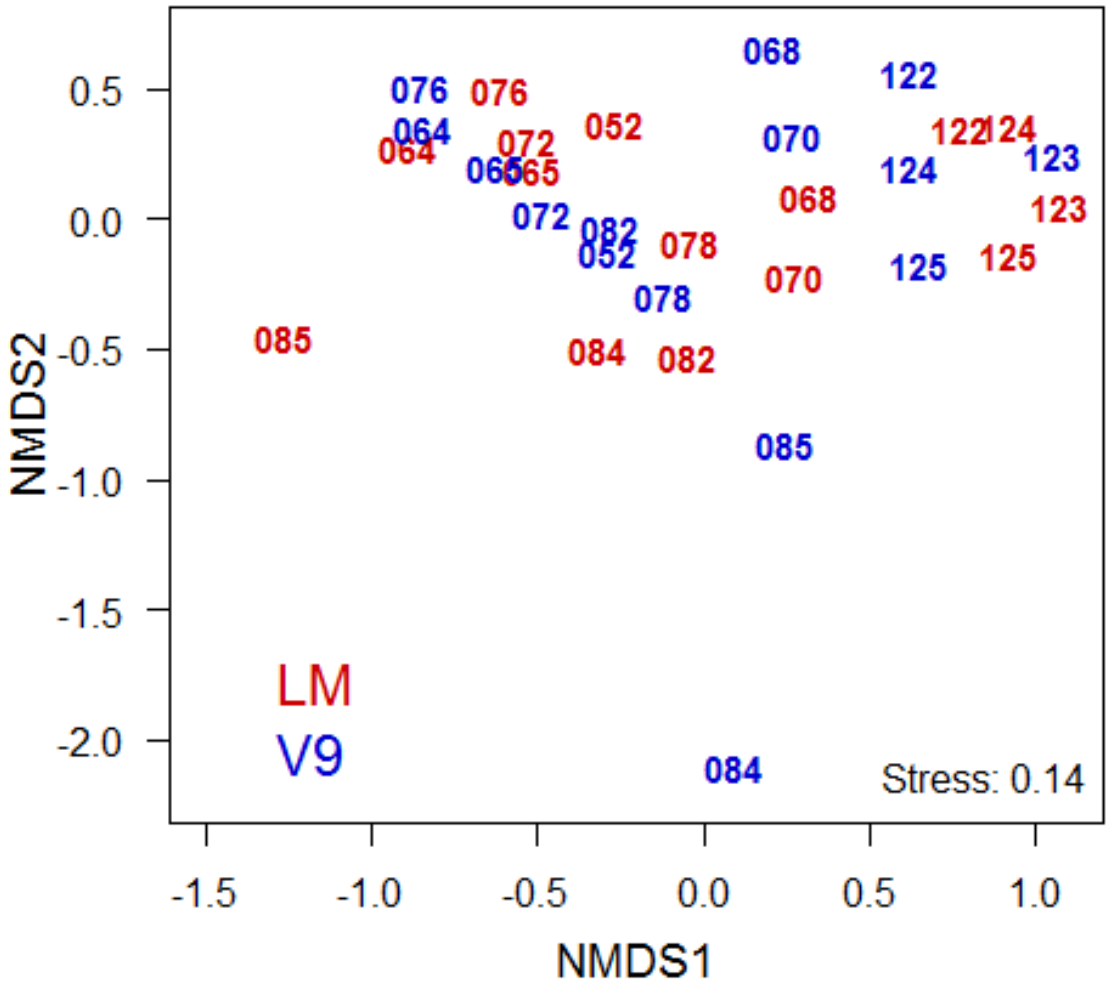
Supplementary figure 2



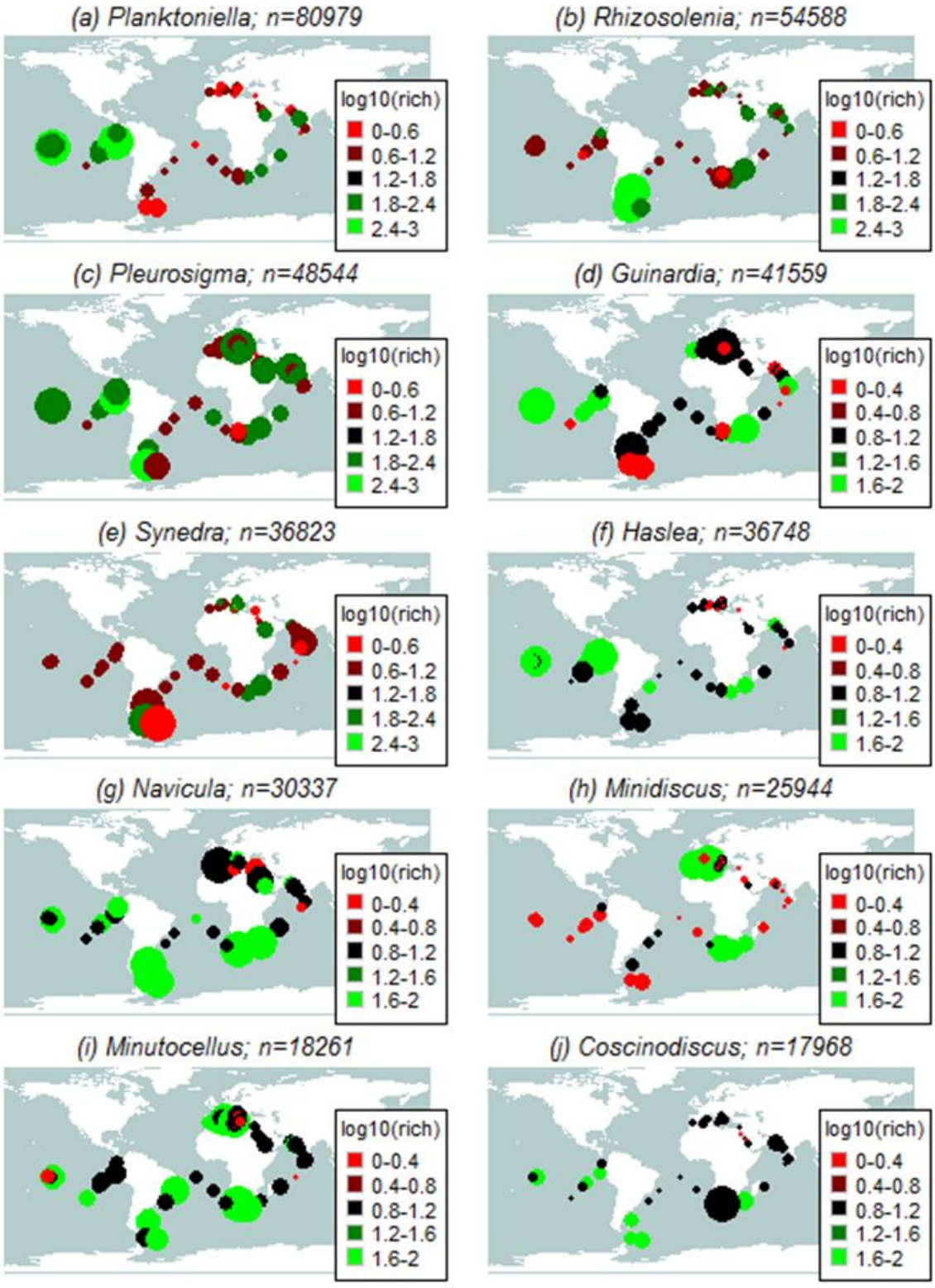
Supplementary figure 3



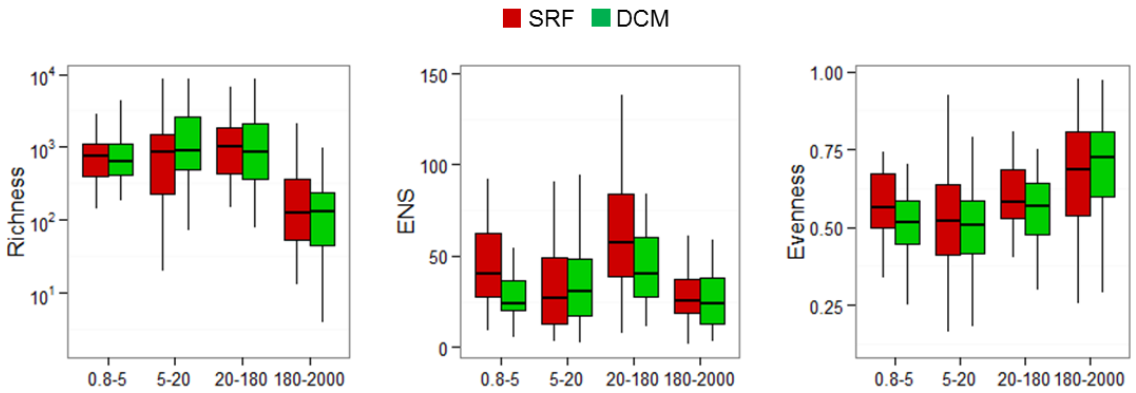
Supplementary figure 4



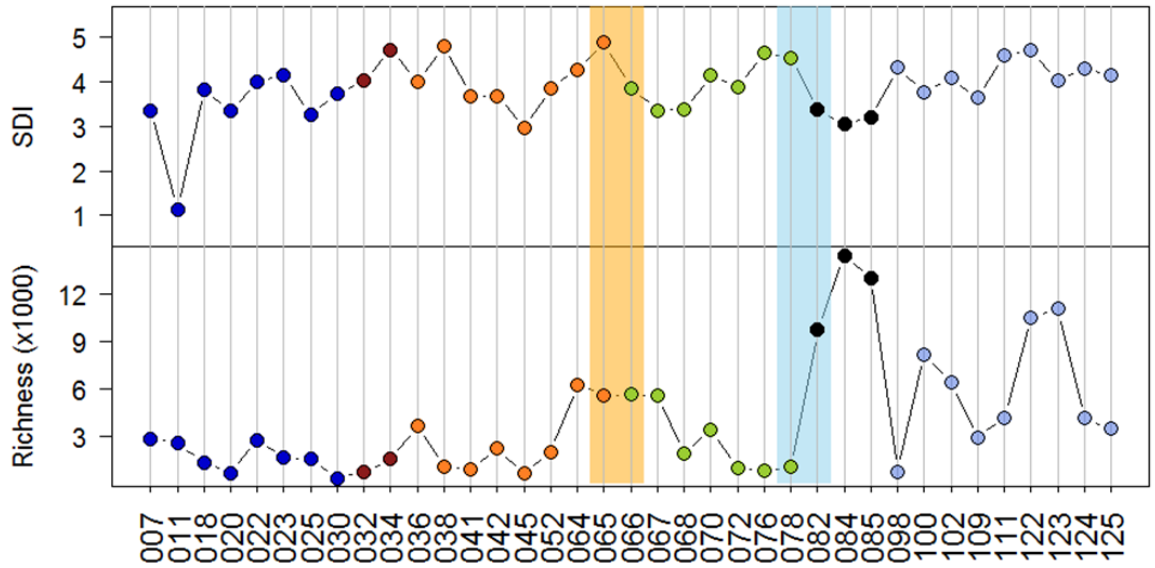
Supplementary figure 5



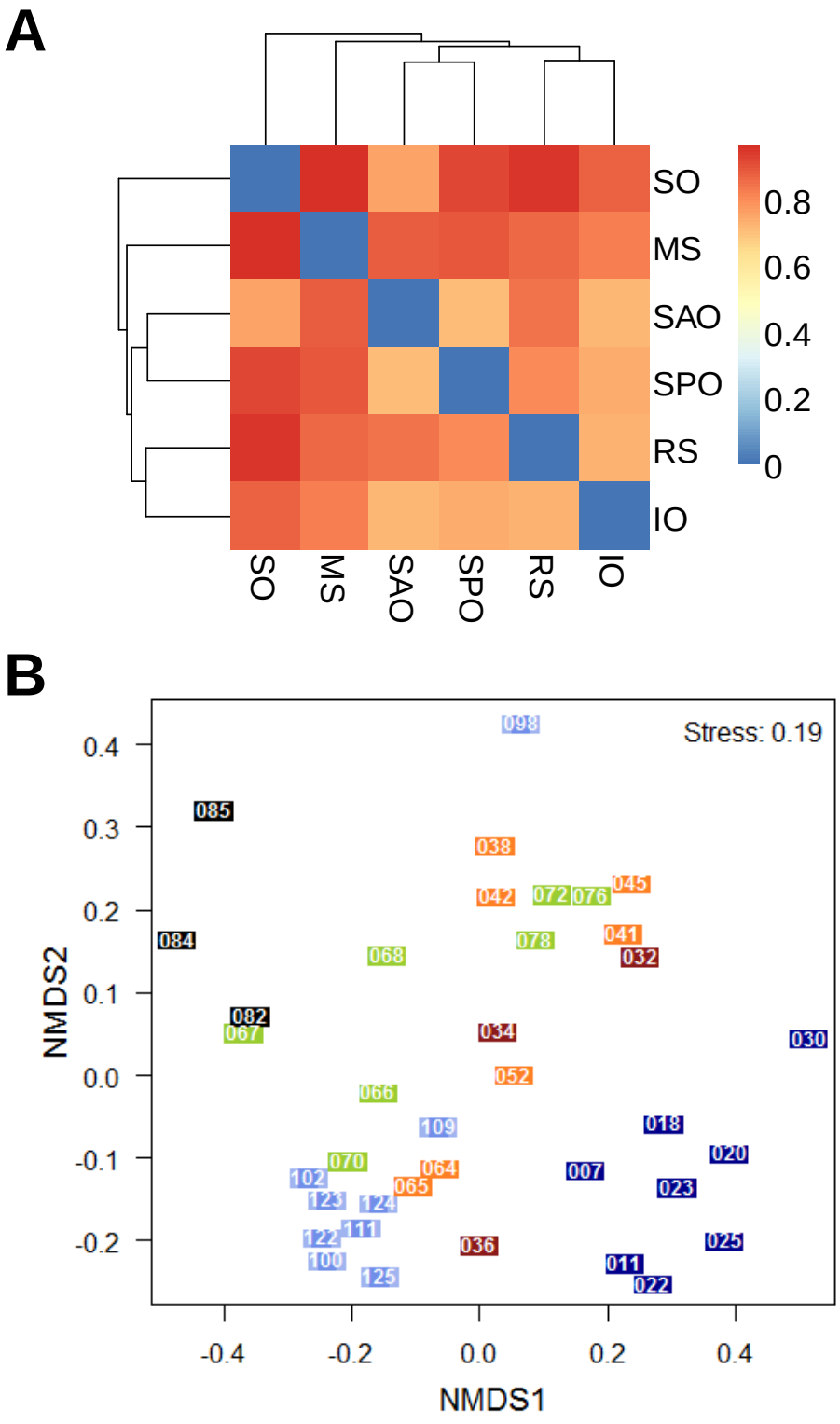
Supplementary figure 6



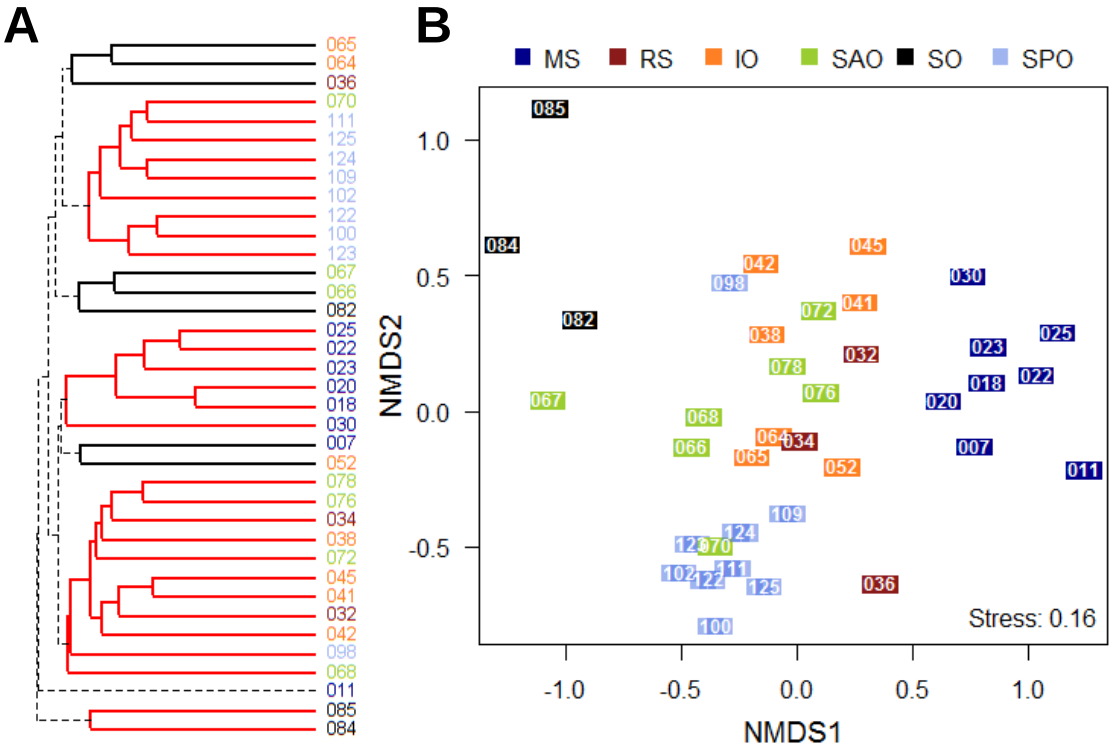
Supplementary figure 7



Supplementary figure 8



Supplementary figure 9



Supplementary figure 10

CHAPTER 3

Niche-based and Spatial Processes Shaping Diatom Community Structure

Summary

Abstract

In the past two decades, many studies have focused on assessing the impact of environmental (niche-based approach) and spatial (neutral processes) drivers in explaining the differences in species richness and composition patterns. In the present study, I was interested to examine these patterns in oceanic diatom communities using a comprehensive set of 18S ribotypes derived from the Tara Oceans sampling expedition. The effects of various abiotic and biotic environmental variables were assessed for each size-fractionated sub-community using Mantel test and canonical redundancy analysis (RDA) and their partial form to control for effects of spatial variables. The models revealed the minimal sets of variables (selected using a forward selection approach) that best explain the variance of ribotype distribution and diversity. Temperature and salinity were found to be the most influential parameters in explaining diatom ribotype composition and richness patterns. The key variables identified were used for variation partitioning and the results demonstrated that the majority (56-81%) of the variation could not be explained by neither measured environmental factors nor spatial distances. This might be accounted for by biological interactions, historical events, ecosystem productivity, and other factors that have not been considered. The results suggest that both niche assembly and neutral processes have significant influences on diatom distributions and diversity. However, it is concluded that dispersal limitation is more important in shaping diatom community structure than environmental heterogeneity.

3.1. Introduction

> *"The niche-assembly perspective asserts that ecological communities are limited membership assemblages of species that coexist at equilibrium under strict niche partitioning of resources. […] The dispersal-assembly perspective asserts that ecological communities are open, continuously changing, non-equilibrium assemblages of species whose presence, absence, and relative abundance are governed by random speciation and dispersal, ecological drift, and extinction."* *- Hubbell 2001 p. 29*

The study of beta diversity patterns is one of the most fundamental issues in biogeography, ecology and conservation. Whittaker (1960) defined beta diversity as "*the extent of change of community composition, or degree of community differentiation, in relation to a complex gradient of environment, or a pattern of environments*". Since then, the term has been used to refer to a variety of phenomena, although all of these encompass some kind of compositional heterogeneity or differentiation between sites (Tuomisto, 2010; Anderson et al., 2011). A fundamental challenge in ecology is to understand the mechanisms that govern underlying compositional heterogeneity. There exists two distinct families of

theoretical models, “niche” and “neutral”, that are being debated for their merits and relevance for explaining patterns in community structure (McGill 2003; Turnbull et al. 2005; Volkov et al. 2007; Chesson 2000; Hubbell 2001, 2005; Chave et al. 2002; Chave 2004; Tilman 2004; Chase 2005; Gaston & Chown 2005). Several studies have proposed to move beyond this dichotomy between niche and neutral theory to a unified theory that can explain the full range of observed patterns in ecological communities. Apparently, there has been a growing consensus for the co-existence of both these processes (Gravel et al. 2006, Leibold and McPeek 2006). Fisher and Mehta (2014) put forth another interesting view that “there is a transition in diverse ecological communities between a selection-dominated regime (the niche phase) and a drift-dominated regime (the neutral phase)”.

The first law of geography states that *“Everything is related to everything else, but near things are more related than distant things”* (Tobler, 1970). In agreement with this, the distance decay of similarity (DDS) is one of the best recognized and fundamental patterns of biodiversity (Whittaker 1975). It has been defined as the decrease in compositional similarities along increasing geographic distances. This decrease in similarity can be explained either by neutral processes including random, dispersal-related (Hubbell 2001) and niche-related processes (environmental variability) (Cottenie, 2005). Beta diversity can provide useful insights into the importance of these two theories in describing the community structure along an ecological gradient. If one assumes neutral theory to be the sole determinant of community structure, then beta diversity is expected to increase along spatial distance (gradients) and to remain constant along an environmental gradient. In contrast, niche theory will demonstrate the opposite under the assumption that it is the sole determinant. Although such strict cases do not occur in nature, the dichotomy facilitates visualization of the underlying ecological processes. For instance, a high rate of similarity decay can be expected for organisms with low dispersal ability (Qian, 2009; Maloney and Munguia, 2011). With their widely varying abundance, diversity, and worldwide distribution, diatoms provide an excellent opportunity to elucidate the potential effect of dispersal ability on patterns in DDS. In addition to the significant contribution of diatoms to total global primary productivity, they play a crucial role in exporting carbon to the bottom of the ocean via the biological carbon pump, owing to their high ballasting effect (Smetacek et al., 2012). Understanding the forces that drive the patterns in diatom distribution is pivotal to explaining ecosystem functioning. Although some studies have emphasized that diatom community composition is predominantly determined by species sorting by the environment (Finlay and Fenchel, 1999; Finlay, 2002; Cermeno and Falkowski, 2009; Cermeno et al., 2010), there is an ongoing debate about the dispersal ability of diatoms (Chust et al., 2013). Gothe and co-workers (2013) performed a small-scale experimental study with diatoms and suggested that their communities are structured by a combination of local (e.g., environmental filtering and biotic interactions) and regional factors (e.g., dispersal related processes). However, most

of the reported studies are constrained either to known genera or to a restricted area of study, so the mechanisms that govern marine diatom distribution have not been systematically investigated at a global scale. Hence, a global analysis can help to develop a more complete understanding of the potential mechanisms causing and maintaining diatom diversity.

Structure of the study. In **Chapter 2** the distribution and diversity of diatoms were investigated and it was found that diatoms do exhibit biogeographical patterns. The study indicated that diatoms exhibit wide geographical ranges with a majority of ribotypes seen exclusively in the South Pacific and Southern Ocean waters. The global study confirmed that they are most abundant in regions of high productivity and at high latitudes whereas they display a high diversity in stations sampled in oligotrophic areas. The study also revealed a considerable amount of novelty for this planktonic group. This work thus addressed general questions on how diatom distribution and diversity varies spatially. In the present chapter, the analysis was extended to identify the processes (niche and neutral) that are potentially responsible for the underlying biogeographical patterns identified in the previous chapter. Firstly, the smallest set of environmental variables that best explain the variance in diatom communities were identified. These variables were then studied to understand their effects on diatom richness and composition patterns. While we hypothesized that environmental heterogeneity is likely to structure the diatom community, we also expected a significant spatial signature in community structure due to neutral dynamics, as other studies have reported that diatoms have large dispersal abilities (Soininen et al., 2004; Wetzel 2012; Verleyen et al., 2009). We therefore examined the decay in beta diversity along both environmental and spatial gradients. The analysis was undertaken for all four size classes of diatoms to examine whether there is a uniform decline in community similarity or not. A faster decay in similarity with distance was expected for the largest size fractions because smaller diatoms are likely to be transported more efficiently by passive currents. Finally, the relative contribution of the niche-based and spatial processes was evaluated.

3.2. Materials and methods

3.2.1. Dataset

***Diatom data set*.** We obtained 293 diatom communities sampled from 46 *Tara* Oceans stations (two depths and four size classes). A total of 63,371 unique tags were obtained from these samples amounting to a total of ~ 12 million reads. A low-abundance filter was applied and ribotypes that were present with a relative abundance ≥0.0001 in at least one sample were selected. Of the total diatom ribotypes, ~40% met this criterion and were selected for further analysis. Based on size class, the ribotype abundance matrix was divided into four sub-communities (0.8-5 µm, 5-20 µm, 20-180 µm,

180-2000 µm) to investigate whether diatoms of different sizes respond differently to environmental and spatial factors. Each sub-community was represented by Hellinger transformed relative abundance data.

Environmental matrix. An initial environmental exploratory dataset (Database 2, de Vargas et al., 2015) was obtained to assess the multivariate effects of environmental variables on ribotype distribution, along with the distribution of plankton functional types (PFT1: Picoeucaryotes; PFT2: Phytoplankton calcifiers; PFT3: Phytoplankton DMS-producers; PFT4: Mixed-Phytoplankton; PFT5:Phytoplankton silicifiers; PFT6: Heterotroph Strontium sulphate skeleton; PFT7: Heterotroph Calcifiers; PFT8: Heterotroph Siliceous skeleton; PFT9: Parasites; PFT10: Picoheterotrophic eukaryotes (without Bacteria); PFT11: Proto-zooplankton; PFT12: Meso-zooplankton) (Lima-Mendez et al., 2015). A total of 33 variables were categorized into abiotic (14 variables; pressure, temperature, salinity, oxygen, depth (mixed layer depth), depth (maximum fluoroscence), depth (maximum N_2), depth (maximum O_2), depth(minimum O_2), NO_2, NO_2NO_3, silicate, Lyapunov exponent, retention), biotic (17 variables; chlorophyll, colored dissolved organic matter [CDOM], flux, net primary productivity [NPP], bacteria, PFT1-PFT12), and temporal (2 variables; season and phase of season) variables. All the abiotic and biotic variables (but not temporal variables) were log-transformed to improve normality and were standardized to scale effects.

***Spatial matrix*.** To understand the mechanisms generating spatial variation in our dataset of diatom ribotypes, spatial structures were modelled using Moran's Eigenvector Maps (MEM), following a data-driven approach described in Dray et al. (2006). This approach allows a set of spatial descriptors (eigenvectors) to be obtained from station coordinates, a network describing the connection between stations, and a weighting scheme for the connections. A total of 12 significant descriptors (MEM1-MEM12) were obtained and were retained as explanatory spatial variables in subsequent analysis. The procedures were carried out using the package *spacemakeR*.

3.2.2. Statistical analyses

***Distance decay analyses*.** For each sub-community, we used Bray-Curtis (compositional) and Jaccard (richness) indices to calculate pairwise similarities between stations. A common set of surface stations among four size classes were selected for this analysis to keep the similar number of comparisons. A matrix of environmental and spatial distances between stations were computed using Euclidean distance. Subsequently, a distance-based regression was performed on each sub-community, regressing Bray-Curtis/Jaccard community similarity against geographical and environmental distance.

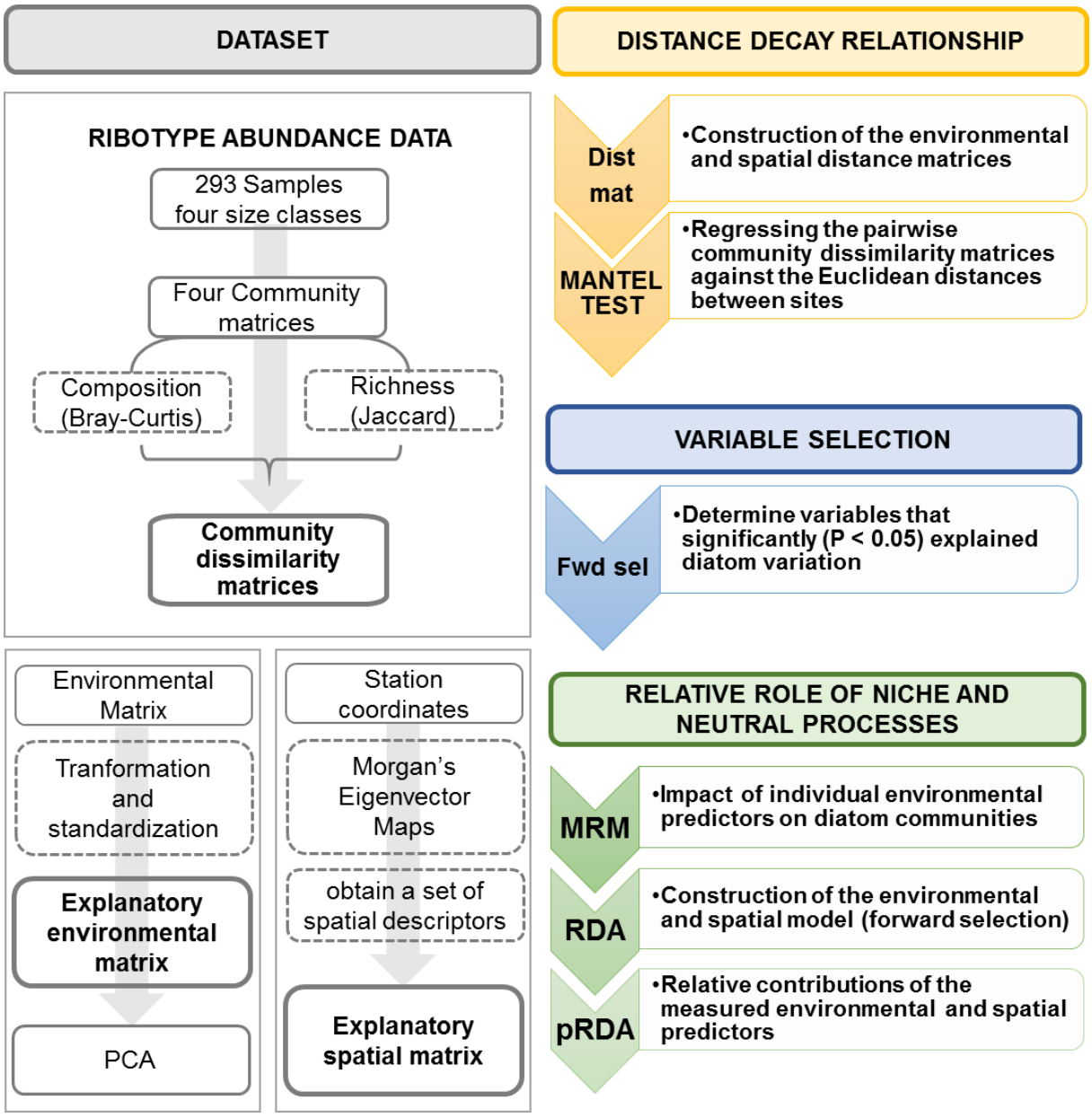
Their standardized regression coefficients were used as a measure of the rate of decay of similarity as a function of geographic distance between stations (Wetzel, 2012). If all the sub-communities exhibited similar distance decay along environmental and spatial distance, then we can deduce that diatoms exhibit similar dispersal ability irrespective of their size.

Mantel tests between pairwise beta diversity and distance. For each sub-community, Mantel analysis (Mantel, 1967) was performed to test for a significant linear distance-decay relationship by computing correlation between spatial distance and community dissimilarity. Mantel analysis using a spatial distance matrix is a direct test of the effect of geographic structure on community composition (Borcard and Legendre, 2002). The influence of environmental variables on community composition was also assessed using Mantel correlations. Finally, we conducted partial Mantel analysis while controlling for geographic distance and vice versa. The associated p-values were estimated using 9999 permutations. The Mantel statistic (r) was reported for each comparison.

Multiple regression analyses on individual environmental variables. To quantify the relative contributions of the individual environmental variables on diatom biogeography, we used permutation-based multiple regression on distance matrices (MRM) as described by Lichstein (2007). This allowed the inferences to be made at the level of individual environmental variables.

Relative roles of niche-based and spatial processes. Previous studies (Lin et al., 2013; Chust et al., 2013) have reported that the total variation in a community can be partitioned into four components, namely (i) variation explained by pure environmental heterogeneity, (ii) variation explained by pure geographic distance, (iii) variation explained by both environmental heterogeneity and distance, and (iv) unexplained variation. Variation partitioning was carried out using a series of partial redundancy analyses (pRDA) to decompose the variance into a pure environmental component, a pure spatial component, a spatially structured environmental component, and residual variation, along with their associated p-values. *Forward selection* (Blanchet et al., 2008) was performed separately on environmental and spatial (MEM vectors) explanatory variables to select a group of most parsimonious variables to avoid an overestimation of the amount of explained variance. Variance partitioning was also used to determine whether the impact of various environmental variables were independent or embedded within each other. The R^2 values were adjusted (Adjusted R^2_a) to account for the number of sampling sites and explanatory variables and the R^2_a statistics were used to generate unbiased estimates of the contribution of the independent variables (Peres-Nato et al., 2006). Monte Carlo permutation tests (9999 permutations) were carried out to compute the significances of the different

Figure 3.1. Summary of study. The workflow illustrates the study designed to determine the role of niche-based



and spatial processes in shaping diatom community structure. *Fwd sel: Forward selection; Dist mat: Distance matrix; RDA: redundancy analysis, pRDA: partial redundancy analysis; MRM: Multiple regression on distance matrices; MEM: Moran's eigenvector maps.*

components. For all statistical analyses, a value of $P < 0.05$ was considered significant. All the data analyses were performed in R (v.2.14.1) using vegan (Oksanen, et al. 2013), ecodist, and MASS packages.

3.3. Results

A total of 11,776,170 reads belonging to 25,766 distinct ribotypes were used in this study. The results are organized into three sections, namely (i) distance-decay along geographic distance, (ii) assessment of which of the environmental variables drive niche-associated differences for each size class, and (iii) relative importance of niche and neutral processes in the assembly of marine diatom communities at a global scale. **Figure 3.1** summarizes the design of the study.

3.3.1. Distance-decay relationship (DDR)

In general, the regression of similarity against geographic and environmental distance demonstrated that community similarities seem to exhibit a significant ($p<0.001$) distance-decay relationship for all size fractions (**Figure 3.2 and 3.3**). The decay in similarity along the spatial gradient suggested that diatoms of the three smallest size classes have nearly similar dispersal ability, which is very different from the largest size-class (decay rate is almost half in comparison to others) (**Figure 3.2**). On the other hand, the decay in community similarity along environmental gradients revealed small differences among different size-classes. The distance decay relationship along environmental distance was strongest for 20-180 µm fraction (slope=-0.0298±0.003, intercept=0.298±0.01, R^2=0.096) compared with the others (0.8-5 µm: slope=-0.024±0.002, intercept=0.262±0.01, R^2=0.08; 5-20 µm: slope=-0.013±0.003, intercept=0.2±0.01; R^2= 0.04; 180-2000 µm: slope=-0.011±0.002, intercept=0.16±0.01, R^2= 0.02) (**Figure 3.3**). Similar patterns were observed for the community matrices computed using Jaccard index of similarity (richness) (**Figure 3.2B and 3.3B)**. These results indicated that environmental heterogeneity may have a differential impact on diatoms of different sizes whereas all the communities, except the largest one, respond in a similar manner to geographic distance. However, a significant relationship between environmental/geographic distance and community (dis)similarity revealed that both play important roles in constraining diatom distribution and composition, similar to that reported for other microorganisms including bacteria (Nekola and White, 1999; Ramette and Tiedie, 2007; Martiny et al., 2011; Lin et al., 2013). However, their influence may be entangled within each other. There was no obvious trends with respect to size (see discussion for detail).

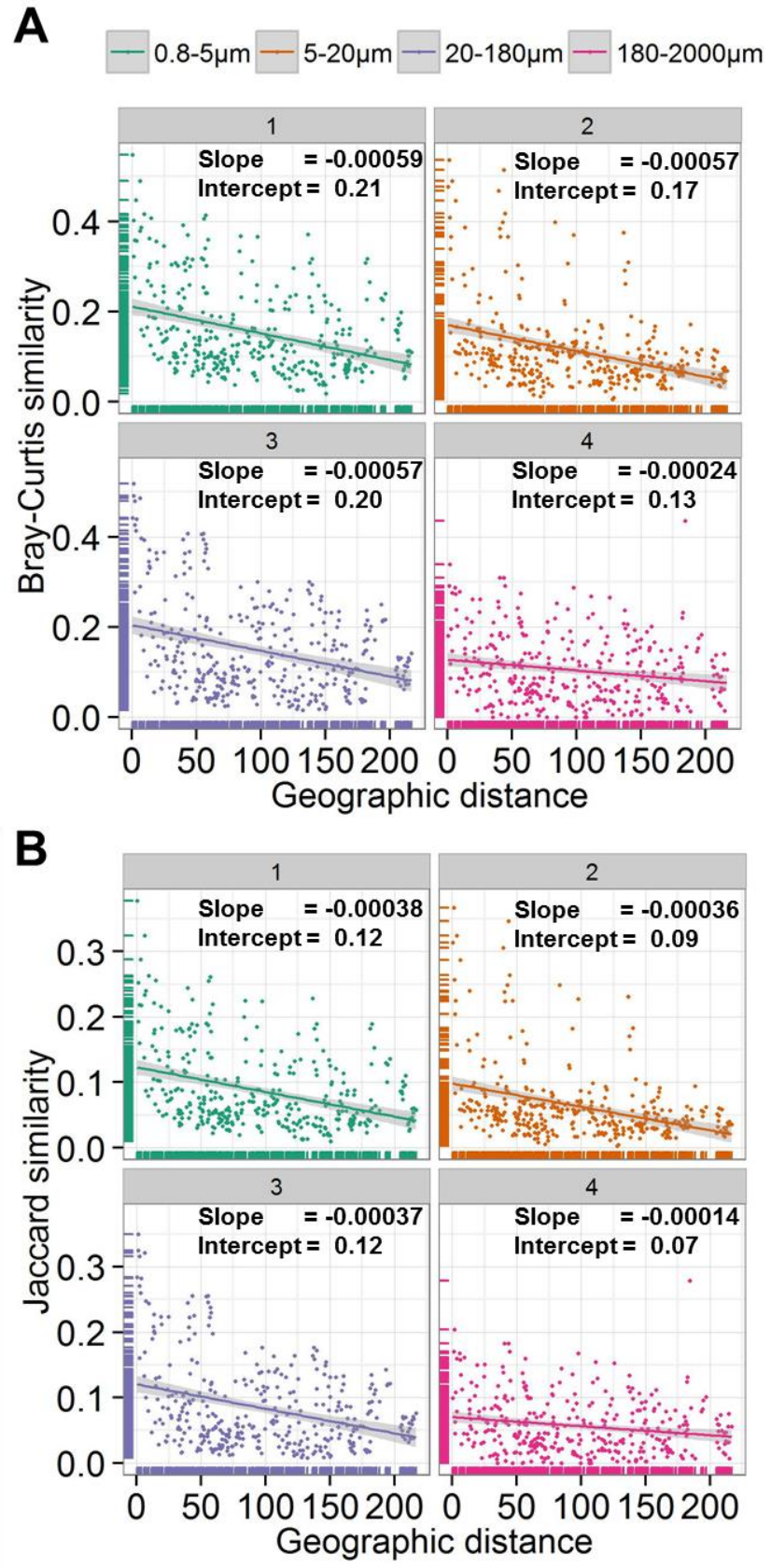


Figure 3.2. Relationship between community similarity and geographic distances. A. Bray-Curtis, and B. Jaccard similarities of diatom communities plotted against geographic distances between sites.

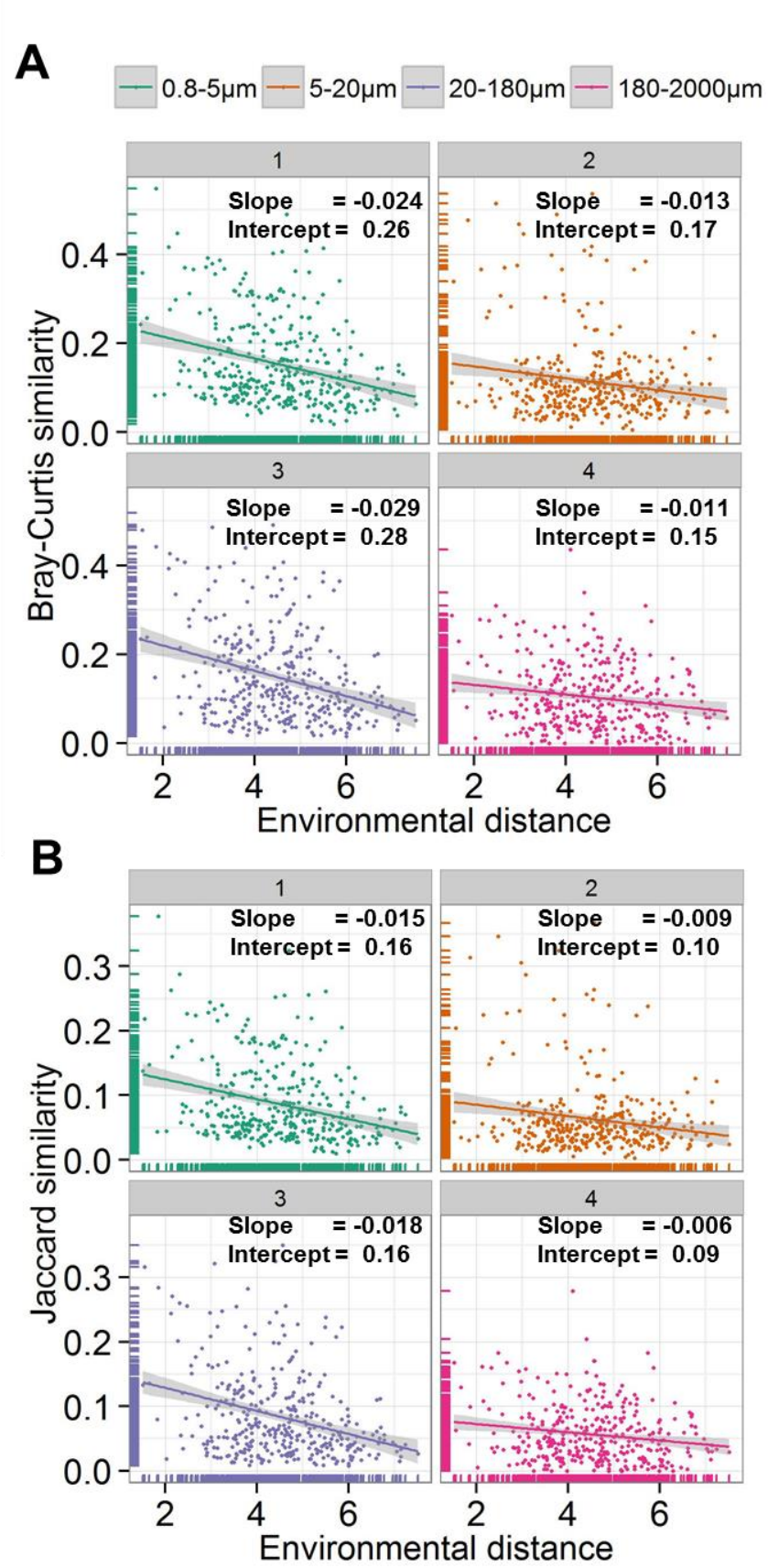


Figure 3.3. Relationship between community similarity and environmental distances. A. Bray-Curtis, and B. Jaccard similarities of diatom communities plotted against environmental distances between sites.

Table 3.1. Mantel's matrix correlation statistics. Mantel tests in general show a significant relationship. The p-value was estimated using 9999 randomizations.

Filter size	Community Dissimilarity	Distance matrices	Mantel statistic	pValue
0.8-5µm	Bray-Curtis dissimilarity	Environmental distance	0.278	0.0002
		Geographic distance distance	0.328	0.0001
		Environmental distance|Geographic distance	0.252	0.0001
		Geographic distance|Environmental distance	0.307	0.0001
5-20µm		Environmental distance	0.209	0.0004
		Geographic distance	0.284	0.0001
		Environmental distance|Geographic distance	0.184	0.0012
		Geographic distance|Environmental distance	0.266	0.0001
20-180µm		Environmental distance	0.311	0.0001
		Geographic distance	0.295	0.0001
		Environmental distance|Geographic distance	0.286	0.0001
		Geographic distance|Environmental distance	0.269	0.0001
180-2000µm		Environmental distance	0.13	0.0289
		Geographic distance	0.144	0.0002
		Environmental distance|Geographic distance	0.113	0.0498
		Geographic distance|Environmental distance	0.129	0.0014
0.8-5µm	Jaccard dissimilarity	Environmental distance	0.281	0.0001
		Geographic distance	0.329	0.0001
		Environmental distance|Geographic distance	0.254	0.0001
		Geographic distance|Environmental distance	0.307	0.0001
5-20µm		Environmental distance	0.219	0.0003
		Geographic distance	0.285	0.0001
		Environmental distance|Geographic distance	0.194	0.0004
		Geographic distance|Environmental distance	0.267	0.0001
20-180µm		Environmental distance	0.317	0.0001
		Geographic distance	0.309	0.0001
		Environmental distance|Geographic distance	0.293	0.0001
		Geographic distance|Environmental distance	0.283	0.0001
180-2000µm		Environmental distance	0.134	0.0211
		Geographic distance	0.15	0.0004
		Environmental distance|Geographic distance	0.117	0.0395
		Geographic distance|Environmental distance	0.135	0.0006

3.3.2. Mantel analysis

Mantel analysis was performed to assess whether community composition was significantly correlated with the environmental and spatial variables. For all sub-communities, the results indicated that community dissimilarity was significantly correlated with spatial distance. This analysis yielded strongly significant relationships (p-value<0.05), demonstrating that stations closer to each other are compositionally more similar to each other (**Table 3.1**). Mantel analysis also indicated that environment had a significant (p-value<0.05) effect on the community composition suggesting that stations that are similar environmentally tend to be similar compositionally (**Table 3.1**). Overall patterns of significance were similar for both composition (Bray-Curtis) and richness-based (Jaccard) beta-diversity metrics. Partial Mantel test demonstrated that both environmental and spatial distances were often entangled with each other. All the sub-communities remained significantly correlated with both the environmental and spatial distance while correcting for the other. Altogether, these results emphasize that diatom community structure is an outcome of both environmental heterogeneity and geographic distance, indicating the role of both niche-based and spatial processes in diatom biogeography, except for 20-180 µm, the others were more related to geographic distance.

3.3.3. Multiple regression analyses on individual environmental variables

To interpret the impact of individual environmental predictors on diatom communities, permutation-based multiple regression on distance matrices was performed as reported by Lichstein (2007). Temperature was found to be the most significant parameter to explain the diatom biogeographical patterns for all size fractions, except 5-20 µm (**Figure 3.4**). In the 0.8-5 µm sub-community, temperature and salinity were found to be the most influential parameters for explaining ribotype composition. Salinity was the variable with the highest impact on the 5-20 µm sub-community, while temperature alone could explain the variance in the 180-2000 µm sub-community. However, in the 20-180 µm sub-community, temperature, salinity, NO_2NO_3, NPP, and phase of the season all had significant and higher impacts. Out of twelve PFTs used in the analysis, only seven were found to be significantly correlated, e.g., PFT1 (Picoeucaryotes), PFT2 (Phytoplankton calcifiers), PFT5 (Phytoplankton silicifiers), PFT6 (Heterotroph Strontium sulphate skeleton), PFT7 (Heterotroph calcifiers), PFT10: (Picoheterotrophs eukaryotes), PFT12 (Meso-zooplankton). Among those, PFT1 and 2 majorly explained the distribution of the smallest size class (0.8-5 µm), PFT2 significantly explained the distribution of 5-20 and 20-180 µm sizes, and PFT5 controlled the distribution of the 180-2000 size-class. Despite being significant, the other PFTs explained only a small percentage of variation. Similar

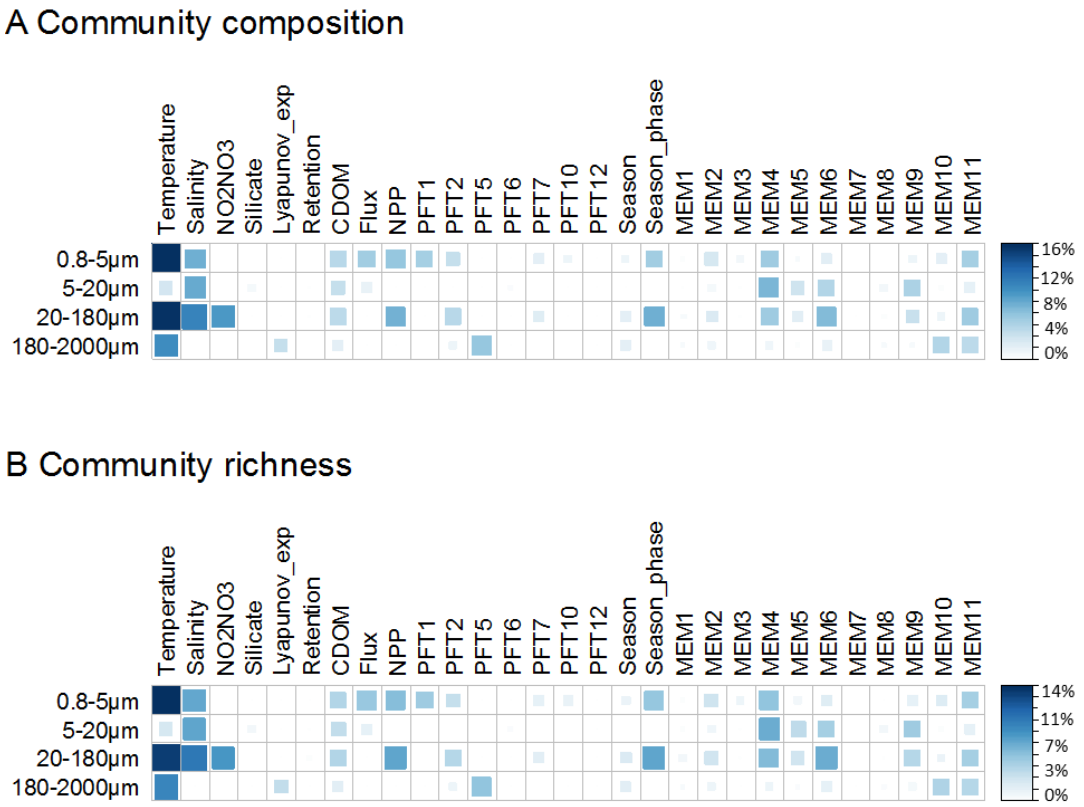


Figure 3.4. Permutation-based multiple regression on distance matrices (MRM). Environmental variables significantly contributing to the variation in diatom community similarity are shown. Each environmental variable was used as an independent matrix. R-squared regression coefficient of each environmental variable is expressed as percent.

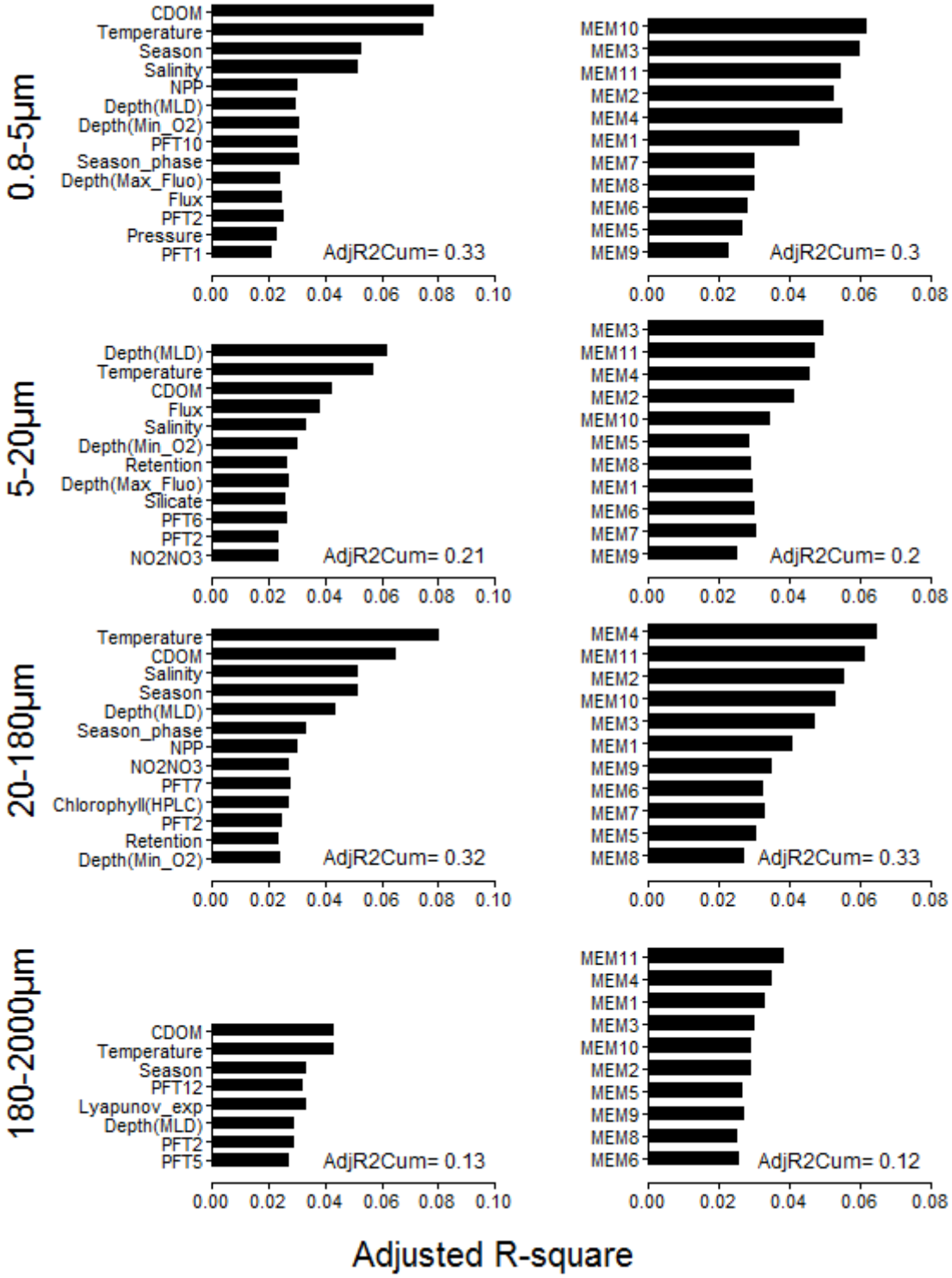


Figure 3.5. Environmental variables selected for each size class. Those environmental variables that significantly contributed to the variation in diatom community similarity were selected using forward selection. The cumulative adjusted R-squared regression coefficients of each environmental variable is shown in the bottom of each panel (significance level, <0.05, 9999 permutations).

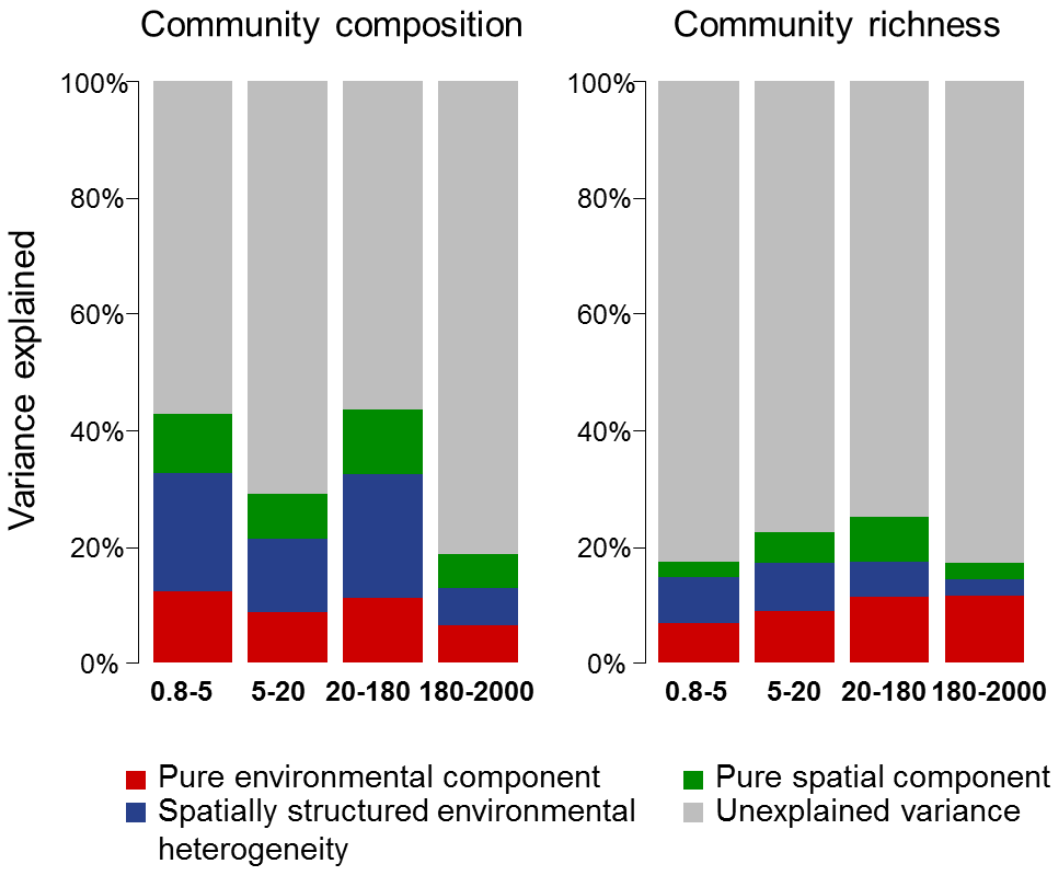


Figure 3.6. Variation in community composition and richness explained by environmental and spatial variables and their shared effects. Barplot showing the results of the variation partitioning procedure carried out on the forward selected environmental and spatial (Moran's eigenvector maps) variables for community composition. For community composition pattern (left panel), pure environment explained 6.5-12% of the total variation. Both environmental heterogeneity and Euclidean geographic distance together explained for 6-20% of the total variation. A pure spatial component accounted for 6-11% of variation in diatom communities. The majority of variance (56-81%) remained unexplained. For community richness patterns (right panel), pure environment explained 7-12% of the total variation. Spatially structured environmental heterogeneity accounted for 2-9% of the total variation. A pure spatial component explained only 3-8% of the total variation in diatom communities. The majority of variance (74-83%) remained unexplained.

variables were seen to explain the variation in richness patterns in each case. The lack of influence of silicate and PFT5 (phytoplankton silicifiers) was highly unexpected.

3.3.4. Relative role of niche-based and spatial processes

For each size-fractionated sub-community, a group of most parsimonious environmental and spatial variables that best explain diatom community composition were obtained using forward selection. For the 0.8-5 µm size fraction, the parsimonious model with fifteen selected environmental variables (p-value=0.005) explained 33.4% of variation (Adjusted R^2 = 0.33) (**Figure 3.5**). The partial model (p-value=0.005) explained 30.8% of variance (Adjusted R^2 = 0.3) when controlling for spatial structure. In the 5-20 µm sub-community, the model (p-value=0.005) retained ten variables that explained 21.1% of the variance (Adjusted R^2 = 0.21) and 20.4% (Adjusted R^2 = 0.2) by partial model controlling for spatial structure (p-value=0.05). In the two larger size classes, i.e., 20-180 µm and 180-2000 µm, the models (p-value=0.005) retained thirteen and eight variables explaining 33.1% (Adjusted R^2 = 0.32) and 13.2% (Adjusted R^2 = 0.13) variation, respectively (**Figure 3.5**). Here, the spatial models (p-value =0.005) explained 33.1% (Adjusted R^2 = 0.33) and 12.9% (Adjusted R^2 = 0.12) variance, respectively.

Variation partitioning of diatom community composition into environmental and spatial components showed that environmental parameters alone could explain only 6.5-12% of the total variation. Both environmental heterogeneity and geographic distance together could explain an additional 6-20% of the total variation (**Figure 3.6**). A pure spatial component alone could only explain a minor portion of the variation present in the diatom communities (6-11%), leaving a major portion (56-81%) of the total variation unexplained. These results show that even though both niche adaptation and neutral processes have significant influences on diatom distributions, neither are able to explain fully the variation in diatom community composition and richness. Biological interactions, historical events, ecosystem productivity, and other environmental factors may account for the unexplained. These results suggest that dispersal ability or currents are more important in shaping diatom community structure than environmental differentiation (i.e., niche).

3.4. Discussion

In the current study I found that the beta-diversity of diatom communities in the *Tara* Oceans data set follows both environmental and geographic gradients, indicating that stations which are similar environmentally or are closer to each other are more similar in their composition. The results further demonstrated that diatom populations do not distribute randomly on a large spatial scale and that their community structure is controlled by a combination of niche-based and neutral processes. This

was in close accordance with published reports and with Hubbell's unified neutral theory, within which he anticipated (2001, p319):

> *"[…] a truly unified theory that at a more fundamental level reconciles these two apparently conflicting perspectives", (i.e., the niche-assembly and the dispersal-assembly approaches).*

However, environmental heterogeneity (pure and spatially structured) was found to explain the majority of the variation in each diatom sub-community. Many studies have reported that diatom species have particular habitat preferences and tolerances; their populations reflecting their response to stress and environmental changes (Schindler, 1987). Recently, several studies with phytoplankton have made similar observations [Pueyo et al., 2006; Chust et al., 2013), concluding that both neutral and non-neutral mechanisms co-occur. In another study, Vellend (2010) recognized four processes that influence community structure, specifically selection, drift, dispersal and speciation. The selection here is analogous to 'niche selection' or 'species sorting', which represents a process where the environment determines the species distribution whereas the other three processes have been accepted as components of 'neutral' processes. The 'niche' theory recognizes environmental filtering and biotic interactions as the major determinants of patterns of species diversity and composition, whereas 'neutral' theory emphasizes the role of dispersal, speciation and ecological drift in explaining the patterns. However, the results presented are contradictory to the views that emerge from recent global-scale studies of fossil diatom assemblages where marine diatom species are reported to possess global dispersal ranges (Cermeño and Falkowski, 2009). This previous study indicated that diatom distribution across oceans does not show any evidence of dispersal limitation and that environmental selection, rather than dispersal, dominates diatom community structure and explains these patterns. In a further study, they confirmed that diatom biogeographic patterns were associated with sharp environmental gradients (Cermeño et al., 2010)

The classical view that "*everything is everywhere but the environment selects*", as summarized by the Baas-Becking and Beijerink hypothesis (1934) outlines a scenario where environmental filtering is the main ecological process governing the different distributions of microbial populations, assuming equivalence among interacting individuals. Under this hypothesis microbial populations are considered as unlimited dispersers and only their niche's differences shape the community composition in a given site. On the other hand, since the development of the Unified Neutral Theory of Biodiversity and Biogeography (UNTB) (Hubbell 2008), dispersal limitation is being considered as an important ecological process capable of reproducing some of the most universal patterns observed in natural

communities. This theory assumes equivalence among interacting individuals and purports that the community composition, at a given site, is shaped by three ecological processes (i.e., birth/death, speciation and dispersal). The previously mentioned niche vs. neutrality debate points out the need for disentangling the relative contributions of different ecological processes (mainly dispersal limitation and environmental filtering) in the assembly of marine diatom communities. But one has to be careful while interpreting the fitted neutral model as it does not necessarily imply the existence of a neutral process behind the pattern, but it might be offering only the simplest explanation consistent with current data.

Dispersal limitation in controlling community structure. Spatial processes have been shown to play a stronger role in several studies with a range of different organismal groups, for instance, ectomycorrhizal fungi (Bahram et al., 2012), bacteria (Lindström and Östman, 2011), and thermophilic archaea (Whitaker et al., 2003). On the contrary, Finlay's "no limits to dispersal" view states that below 1 mm body size 'everything is everywhere, but the environment selects' (Finlay, 2002). Other studies have demonstrated that dispersal limitation varies between different phytoplankton groups (Chust et al., 2013). They further demonstrated that coccolithophores have a higher spatial structuring than diatoms and a higher slope in distance decay, indicating a dispersal limitation. Earlier, Nekola and White (1999) also reported that higher dispersal ability causes a decrease in distance decay rates. In other words, selection and drift increase the strength of the distance–decay relationship (i.e., they steepen the slope) whereas dispersal weakens the distance–decay relationship (i.e., it flattens the slope). Hanson et al. (2012) reported that evidence for biogeographic patterns can be divided into two categories that are "endemism" and a "distance-decay relationship". My previous results (**Chapter 2**) demonstrated the first aspect, i.e., that diatoms exhibit little overlap in different sampling sites and oceanic provinces, indicating endemism at local and regional scales. This study demonstrated that community similarity decayed significantly along geographic distance, although there was no correlation between the distance decay rate and diatom size. However, the presence of distance-decay relationship in diatoms, albeit weaker, emphasizes the importance of dispersal in diatom community assembly. In general, the results were unable to support our initial hypothesis that dispersal ability is inversely related to size (**Figure 3.2** and **3.3**). An unexpected outcome was the smallest decay in similarity (absolute slope) in the largest size class (180-2000 µm) along both the environmental and spatial gradients. The non-uniform decline observed in community similarity suggests that some diatoms are more limited in dispersal than others. One may expect that niche-related processes are more important in structuring local communities for organisms with high dispersal ability (180-2000 µm), as they can reach favorable locations in comparison to the ones with limited dispersal ability

(Martiny et al., 2006). In such organisms, species sorting may drive composition based on their environmental requirements.

As mentioned above, diatom communities of different size-classes exhibit very little differences in their distance-decay rates. This may be due to the notion that dispersal ability is a species trait. In one previous study, Wetzel et al. (2012) tested whether the rates of decay in community similarity differ between diatom growth forms with different dispersal abilities (periphyton with lower dispersal ability and plankton with higher dispersal ability). They found that the rates of distance-decay in community similarity were higher for periphyton than for phytoplankton indicating the lower dispersal ability of periphytic taxa. In my study, the same diatom can be present in more than one size fraction and it can be expected that a community with nearly identical composition in terms of species will exhibit similar decay, as observed (very little difference in the slopes). Further, the studies that have reported dispersal ability to be inversely related to body size were performed on different biological groups (De Bie et al., 2012; Farjalla et al., 2012; Shurin et al., 2012; Van der Gucht et al., 2012). Therefore, one may expect a very small (or no) difference among the dispersal abilities among individuals of different sizes of the same biological group. Although the difference in the rate was not very pronounced, a comparison of slopes revealed that the largest communities (180-2000 µm) have the highest dispersal ability (less steep) in comparison to their smaller counterparts. This may be due to the ability to attach to other larger organisms. These attached larger diatoms may exhibit a different spatial structure than metacommunities of weakly attached smaller diatoms. Moreover, by virtue of the durability of their siliceous shells, diatoms achieve a wider geographical distribution; not only through wide dispersal by ocean currents but also with the help of larger organisms, such as tintinnids and other zooplankton. This can be majorly attributed to their resistant shell and their ability to withstand long periods of desiccation. With the onset of favorable conditions, larger diatoms divide more frequently and are subsequently removed from the system through dispersal. Thus, cell division and cell death influence net turnover and hence, their dispersal. In addition, this is essentially a sub-tropical expedition and thus, (i) diatoms can be found in the form of mats (e.g., *Rhizosolenia*) and hence in large size fractions, beyond expectation; (ii) if the larger size is made of such oligotrophic species, the higher dispersal ability of this size class might reflect the larger size of oligotrophic gyres when compared to eutrophic regions (e.g., upwelling or specific currents such as the ACC).

Role of environmental heterogeneity in controlling community structure. The present study has provided a unique opportunity to investigate the spatial distribution of diatoms across a large geographical scale covering seven oceanic provinces and to examine relationships between diatoms and environmental variables. Temperature was the most significant environmental factor influencing

diatom composition and richness patterns for all diatom size-classes. Our results are well in concert with previous reports on temperature as the key variable associated with planktonic diatom communities (e.g. Patrick, 1971; Yao et al., 2011). Salinity was another variable that exhibited a major influence on diatom community structure. The relationship with environmental parameters were found to be intricate and strongly size-dependent. The case of 5-20 µm showed augmented correlation with salinity, which thereby suggests a relation with geography (e.g. MS and RS) and seems to be an interesting topic for investigation for the future. One fundamental issue in explaining the community structure is that many of the environmental variables are inter-correlated among themselves, and so it may be difficult to attribute a causal mechanism to a single variable even if it is significantly correlated. Working on the *Tara* Oceans Mitags dataset for prokaryotes, Sunagawa et al. (2015) demonstrated that the effect of highly correlated environmental predictor variables (temperature and oxygen) could be disentangled by independently modelling associations (using a binary input matrix to the elastic net fitting routines) of each of the two predictor variables with taxonomic/functional composition for the SUR samples. They then tested the strength of these associations in DCM layers, where correlations between the two factors were much weaker, which allowed them to effectively decouple dissolved oxygen from temperature. A similar approach could be applied to this dataset to disentangle the effect of highly correlated factors.

Mapping diatoms onto the metacommunity paradigms. A metacommunity has been described as "*a set of local communities that are linked by dispersal of multiple interacting species"* (Wilson 1992). Leibold et al. (2004) identified four "metacommunity paradigms":

Type A: "*the neutral view"* which assumes ecological equivalence of species with dispersal-limited communities (Hubbell 2001; Bell 2001);

Type B: "*the species-sorting view"* which assumes a sufficient dispersal of species in a heterogeneous environment with associated niche differences (Chase & Leibold 2003);

Type C: "*the mass-effects view"* assumes a dispersal through source-sink relations that is independent of resource gradients (Holt 1993; Mouquet & Loreau 2002; 2003);

Type D: "*the patch-dynamic* view*"* which is a form of niche differentiation where trade-offs lead to spatiotemporal niches (Hastings 1980; Tilman 1994).

However, it is unlikely that all the species in a community will uniformly conform exclusively to one of the above mentioned perspectives. On one hand, it is difficult to draw clear cut boundaries among different metacommunity paradigms, while on the other there is a high probability that a real ecological community is subjected to a combination of these (Leibold et al., 2004).

In the present study, the variation in community composition was decomposed into four components (pure environment, pure spatial, shared effect, unexplained). The significance structure of these variation components was used to estimate the most important process (environmental, dispersal, or a combination of both) in determining the community structure, as described by Cottenie (2005). Based on the decision tree proposed by Cottenie (2005), we found that the diatom metacommunity is subjected to 'species-sorting + mass-effect' which corresponds to a combination of types 'B' and 'C' proposed by Leibold (2004). As mentioned above, this type of metacommunity is structured by both environmental and spatial variables, independently of each other.

In particular, I studied the dispersal behavior of diatoms with ribotypes as the units of dispersal, by applying the "terrestrial" ecology approaches to the marine ecosystem. Previous studies were only based on geographically restricted areas and, hence, the environmental optima and tolerances could not represent the full ecological range of the studied taxon, as noted by Pienitz et al. (1995) and Weckstrom and Korhola (2001). Therefore, it is appropriate at this level to employ metacommunity theories or other related approaches mainly developed for terrestrial ecological studies, for obtaining useful insights for diatoms as also suggested by Grimm et al (2003). The lessons that would be learnt while verifying these hypotheses, will help to achieve crucial understanding about the key processes and structures dominating the marine environment. Nonetheless, these "terrestrial" ideas will likely pose limitations in applications for the marine system. The major shortcoming would be to identify a metacommunity. Another limitation of such a study would be the biased geographical distances (as the accurate distance in the ocean is based on Lagrangian measurements i.e. the one that follows the current). For instance, Stations 78 and 82 appear close on the map (**Figure 2.1A**), but they are much farther. In milieu of these suggestions, the results provide a baseline for further studies.

Life strategies. Despite many efforts in the past, our understanding of how diatoms respond to environmental changes is still incomplete and limited. A current view is that diatoms are r-strategists, i.e., upon the onset of favorable conditions (e.g., light and nutrients) their rapid turnover allows them to dominate the phytoplankton community, associated with mixed waters and unpredictable conditions. Margalef (1978) examined the selection processes that organize different life-forms of phytoplankton as alternative strategies for survival. This study emphasized that the significant importance of an environmental factor in explaining a community structure has a genetic basis which may shift or evolve. This can be even true for any closely related organism when responding to a variable that is significantly correlated, e.g., temperature or salinity. This study also put forth the notion that general nutrient availability together with physical environment, advection and turbulence controls phytoplankton community composition (Margalef, 1978). These results along with those

presented in **Chapter 2**, regarding the distribution and diversity of diatoms, suggests that they are not exclusively r-strategists as previously believed but can exhibit varied life strategies. Their dispersal abilities vary a lot between species, thereby making it difficult to generalize as it is constrained by the match between optimality of life strategies with the environmental conditions along the ocean currents. In addition, it can be associated with a specific biology (e.g., adaptation to oligotrophy) rather than with their size. This is in consensus with Finlay's "no limits to dispersal" view (Finlay, 2002), which suggests that diatom community composition is predominantly determined by species sorting (Finlay and Fenchel, 1999; Finlay, 2002). Nevertheless, size is a dominant trait for diatoms that armors them with several capabilities (e.g., reproduction, nutrient uptake, grazing defense).

In general, marine ecosystems are believed to be highly connected and prone to frequent environmental changes (natural or anthropogenic). Many studies have demonstrated that if dispersal rates are low in comparison to the frequency of environmental changes, they primarily regulate the "assembly history" of local sites by dictating which species are present at that site (Post and Pimm 1983; Rummel and Roughgarden 1985; Drake 1990; Morton and Law 1997). The presence of species will thus depend on their ability to colonize that site following the onset of favorable conditions. On the other hand, a high dispersal rate will alter the local population. As the newly immigrant population is not self-sustaining, its presence will lead to competitive suppression of other local self-sustaining populations which sometimes may get driven to extinction (Amarasekare and Nisbet 2001; Mouquet and Loreau, 2002; Leibold and Norberg, 2004). An important property of diatoms (and many phytoplankton in general) are the resting stages that represent organisms that are ecologically active but are protected from disturbances. Another notion worth mentioning is the constraints in population imposed by advection. It acts in two ways, firstly if the species transported via advection to an area that is not suited for them, which may lead to mortality. Secondly, these newly migrated individuals may well utilize the resource, reproduce and serve as food to consumers in potentially stronger ways. To conclude, it can be hypothesized that the variation in diatom distribution and diversity can be linked to the connection between different water masses and may not be the result of randomly distributed entities. Also, niche adaptation has a significant influence in governing diatom distribution. Biological interactions such as predation, parasitism and symbiosis are forces likely to be additional factors influencing the structures observed. This study also presented "sub-networks" or sub-ecosystems associated to diatoms and demonstrated grazers as a possible cause of the abundance modulation. The current results demonstrate interesting novelties, e.g. phytoplanktonic calcifiers correlated strongly to all the size-classes, which deems an in-depth analysis.

Nonetheless, the patterns illustrated in this study suggest that this may warrant more thought on how processes that occur at larger spatial scales alter dynamics and patterns of variation seen at the local scale. A comparison of distance-decay relationships among taxa with similar ecological requirements (thus controlling its effect) but with different dispersal ability to evaluate the pure effects of dispersal ability on beta-diversity patterns in diatoms remains a perspective for future work. The local importance of seasonality is another parameter that could be considered in a future study.

CHAPTER 4

A Metabarcoding-based Assessment of Diatom Assemblages

Summary

Abstract

This study presents preliminary results from a global metabarcoding study dealing with diatom community assemblages. A set of selected "*common diatom ribotypes*" from 46 sampling stations were organized into assemblages based on their distributional co-occurrence. Using Ward's hierarchical clustering, nine clusters were defined. The number of ribotypes and reads varied within each cluster; three clusters (II, VIII and IX) contained only a few reads whereas two of them (I and IV) were highly abundant. Of the nine clusters, seven can be divided into two categories defined by a positive correlation with phosphate and nitrate and a negative correlation with longitude and, the other by a negative correlation with salinity, temperature, latitude and positive correlation with Lyapunov exponent. All the clusters were found to be remarkably dominant in South Pacific Ocean and can be placed into three classes, namely Southern Ocean-South Pacific Ocean clusters (I, II, V, VIII, IX), South Pacific Ocean clusters (IV and VII), and cosmopolitan clusters (III and VI).

4.1. Introduction

"*If we reduce nature to what we understand, we would not be able to survive*."
- Hans-Peter Dürr

Several concepts of species association have been developed since the nineteenth century (Whittaker, 1962). Dale (1977) stated that "Interspecific associations arise when two or more species co-occur either more or less frequently than expected due to chance alone. Positive associations between two species can occur when both select the same habitat or have the same environmental requirements". Legendre and Legendre (1978) gave a statistical definition to species association as "*a recurrent group of co-occurring or correlated species",* without emphasizing whether the association was positive or negative. Instead of having to describe the biology of each species individually, associations provide a means to assign ecological requirements common to most or all species in an association. One of the potential implications of such studies can be to employ these associations to predict environmental characteristics (Legendre, 2005).

Ecological communities are characterized by a certain degree of diversity (Olszewski, 2004) and complex interactions among components operating on different spatial and temporal scales (Storch and Gaston, 2004; Steinnauer, 2009). There has been a varied interest in understanding the relationship between spatial distribution of microorganisms and the local environmental factors controlling their distribution. Association analyses are potentially a valuable tool with which one can generate hypotheses about the factors responsible for the distributional patterns. Understanding of

such patterns may help in the comprehension of underlying evolutionary or ecological processes. The most common method for classifying large numbers of individuals according to their geographical preferences/attributes is cluster analysis. These methods offer identification of groups of significantly associated species on the basis of their distributional co-occurrence; which are rational units within which ecological connections can be scrutinized.

Diatoms are major players in the marine photosynthetic world, and represent diverse, ubiquitous, and sensitive environmental indicators (Round et al., 1990). As their distribution is strongly affected by environmental conditions (Charles, 1985), they have been successfully demonstrated as valuable proxies for a wide range of physico-chemical variables (Blanco et al., 2013; Stoermer and Smol, 1999). Interestingly, they may also be useful for studying the effect of trophic interactions on community structure as reflected by significant grazing effects on them, reported in various experimental studies (Lange et al., 2011; Gothe et al., 2013). However, the degree of their utility is dependent on understanding (i) the effect of environmental variables on their distribution, (ii) the extent to which environment explains variation in their community structure, and (iii) the general underlying processes that generate patterns in their distribution and diversity.

Structure of the study. A previous chapter has revealed global diatom distributions and their remarkable diversity (**Chapter 2**). The results have demonstrated that connectivity of local water masses to ocean circulation has a major impact on marine diatom biogeography. Also, the substantial sharing between sampling stations separated by great distances (for instance, equatorial stations) suggested a widespread but not ubiquitous distribution. Most were characterized by a very different composition of ribotypes, with many of them being exclusively seen only in one province. Despite this, a remarkable number of ribotypes were also found to be shared among a combination of two provinces. Nonetheless, the majority of variation in the diatom communities remained unexplained by niche-based or spatial processes (see **Chapter 3**). This current study was designed to understand diatom community organization by identifying groups of associated ribotypes on the basis of distributional co-occurrences. In this chapter, a set of common diatom ribotypes representing the majority of diatom abundance (see Materials and methods for details) was used to investigate whether co-occurring ribotypes can be significantly associated into recognizable clusters. And if so, do they tend to exhibit a distinct behavior in a way that these clusters can be expressed as a function of varying environmental parameters? Thus, these clusters can be reasonable entities which can be potentially useful in further examining ecological relationships. The major objectives of this study were (i) to identify significant ribotype clusters, (ii) to characterize each cluster taxonomically and by their

distributional patterns, and (iii) to search for environmental determinants which could explain or help delineate these clusters.

4.2. Materials and methods

4.2.1. Study area and dataset

The dataset for this study is derived from the *Tara* Oceans expedition (Karsenti et al., 2011) and represents 293 planktonic samples from diverse oceanic provinces. At each station, plankton communities were obtained for four size fractions from two water-column depths (sub-surface water (SRF) and the Deep-chlorophyll maximum (DCM)). Total nucleic acids (DNA + RNA) were extracted from all samples, and the hyper-variable V9 region of the nuclear 18S rDNA was PCR-amplified (Amaral-Zettler et al., 2009). The V9 reads were quality checked and to reduce the influence of PCR and sequencing errors, only sequences seen in at least two different samples with at least 3 copies were retained, giving a total of ~580 million reads represented by ~2.3 million unique metabarcodes (De Vargas et al., 2015). These unique barcodes were taxonomically assigned to known eukaryotic entities based on the PR2 database (Guillou et al., 2013). From this, metabarcodes assigned to diatoms (at a percentage identity of ≥ 85% to the reference sequence) were selected (**Figure 4.1**). Considering that this dataset contains many rare metabarcodes, only those metabarcodes that appeared in at least ten sampling stations surveyed with ≥100 reads (pooled across all samples) were selected for this study and were designated “common ribotypes”. The matrix was Hellinger transformed prior to analyses.

An environmental matrix consisting of a set of environmental variables describing physico-chemical, nutrient and chlorophyll data of each sample was created (**Table G4.1**). All environmental variables were checked for normality and were log-transformed to reduce skew distributions. To avoid problems of logarithm zeros, the number one was added to the abundance of each ribotype {log10(x+1)}. After that, the transformed data were proportionally scaled between 0 and 1 in the range of the minimum and maximum for each variable. In addition, spatial variables, including longitude and latitude were obtained for each sampling site.

For further characterization, all the stations were organized in six classes based on oceanic provinces, i.e., Mediterranean Sea (MS), Red Sea (RS), Indian Ocean (IO), South Atlantic Ocean (SAO), Southern Ocean (SO), and South Pacific Ocean (SPO).

4.2.2. Statistical analyses

The environmental dataset gathered a total of 37 variables (**Table G4.1**), and most of them were

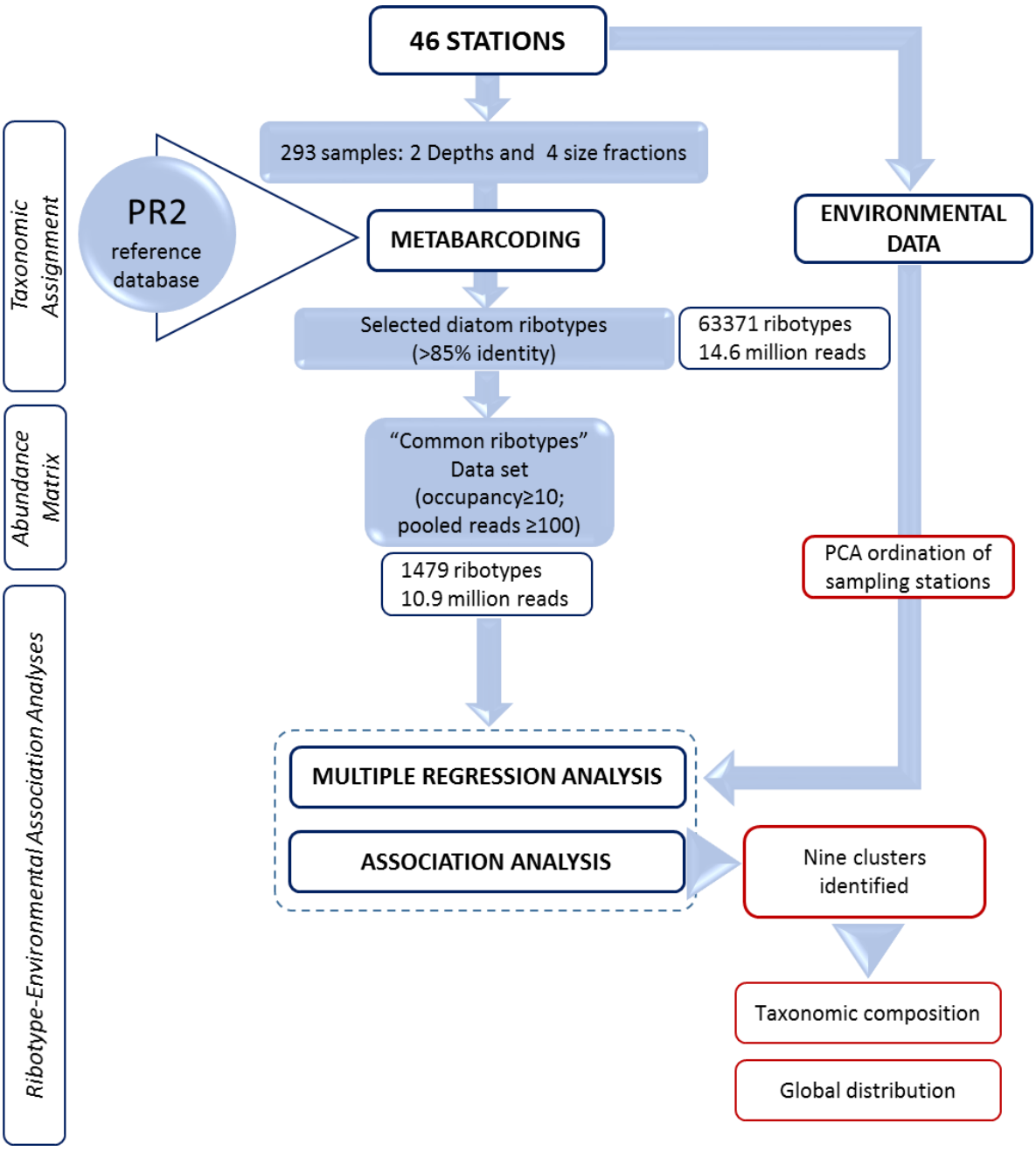


Figure 4.1 Outline of the study.

intercorrelated. For this reason, a variable-reduction procedure was used to identify a subset of predictor variables that minimized multicolinearity and maximized correlation with diatom distributions. This was done in two steps: (1) selecting environmental variables that were significantly correlated to diatom distribution, and (2) analysing the environmental correlation matrix to identify highly correlated variables and then selecting either one or two predictors that maximized the percentage deviance explained in diatom distribution. This second step was done using multiple regression models.

PCA ordination of sampling sites. A principle component analysis (PCA) was employed to examine the variability in environmental parameters among and within the defined oceanic provinces. Analysis of similarities (ANOSIM) was performed to test whether there is a significant environmental difference between stations grouped according to the oceanic provinces.

Influence of environmental variable on each ribotype. Spearman correlation was computed and its significance was used to determine whether any independent environmental variable was significantly correlated to each ribotype. Statistical significance was assessed at a p-value<0.05.

Identifying significant diatom associations. Significant ribotype associations (cluster, hereafter) were identified using the method described by Legendre (2005). Firstly, a Pearson correlation matrix was computed for all the ribotypes and a distance matrix (1-correlation matrix) was obtained. Hierarchical clustering based on Ward's minimum variance method that aims at finding compact, spherical clusters was performed using *hclust* to identify clusters of covarying ribotypes. The vector for ribotype membership was obtained. To identify significant associations, Kendall's coefficient of concordance was computed for each cluster through a permutation test using function *kendall.global* (vegan package). Kendall's coefficient of concordance ranges from 0 to +1. A value closer to zero represents lack of agreement while a value closer to 1 represents perfect agreement in the rankings of the ribotypes among samples. Clusters that were globally significant (p-value < 0.05) were retained for further analysis. Spearman's correlation between each environmental variable and each of the "common ribotypes" was computed to define a common characteristic in each case.

Determining the relative importance of selected environmental variable. To assess their relative importance, R package *relaimpo* (Gromping, 2006) was used. This function calculates how much of the variance in community structure can be explained uniquely by each variable. This analyses was performed separately and in the exact same way for the each of the identified ribotype clusters. All statistical analyses were performed using R 2.5.1 (R Developmental Core Team).

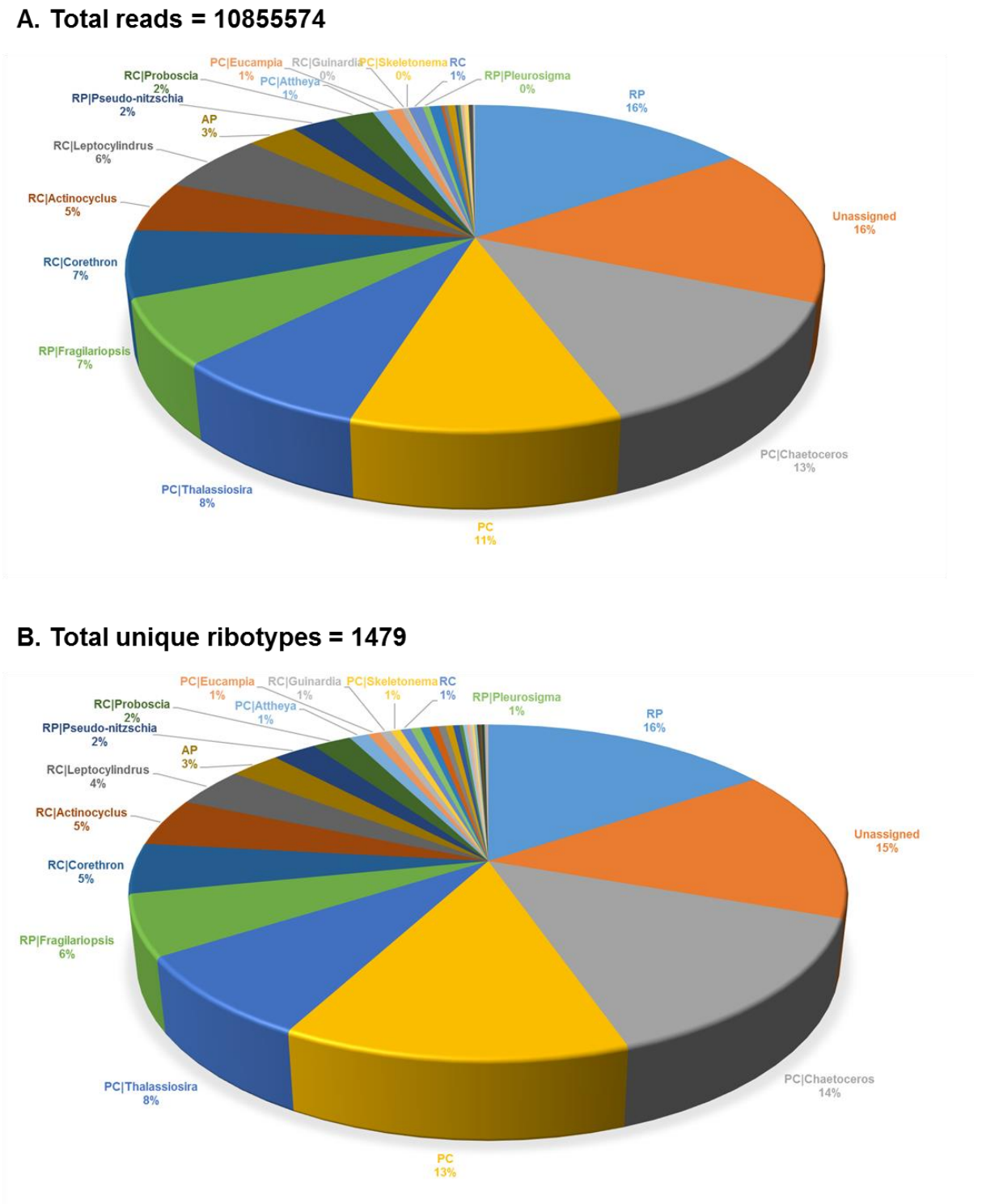


Figure 4.2. Taxonomic composition of the dataset used in the study. Nearly 45% of the reads belonged to either unassigned diatoms or unassigned raphid pennates or unassigned polar centric (for details on taxonomic assignment, please see Chapter 5). (A) Total reads, and (B) Total ribotypes. *AP: Araphid pennate, RP: Raphid pennate, RC: Radial centric; PC: Polar centric; Unassigned: Unassigned diatoms*.

4.3. Results

The "common ribotype" dataset used here consisted of 1479 ribotypes, representing ~90% of the total ribotype abundance (the total set of 63,371 ribotypes corresponds to 12,077,752 reads, while the 1479 ribotypes represent 10,855,574 reads). Of these, 692 ribotypes (46.8%) could not be assigned to either a genus or species. Nearly 45% of the reads belonged to either unassigned diatoms or unassigned raphid pennates or unassigned polar centric diatoms **(Figure 4.2**). The most abundant ribotypes tended to be the most cosmopolitan (**Figure 4.3A)**, while the majority were seen in 10-20 stations and represented the lowest abundance class (100-1000 reads) (**Figure 4.3A and B**). The environmental variables that were found to significantly correlate with diatom distribution are shown in **Figure 4.4A**, and their relationships with each other can be seen in **Figure 4.4B**. From this group of highly correlated environmental variables, one predictor variable that maximized the percentage deviance explained in diatom distribution matrix was selected.

The results presented here are organized into five sections, (i) ordination of environmental variables to examine the variability among and within oceanic provinces, (ii) correlation of individual variables to each ribotype, (iii) identification of significant clusters and describing their characteristics, (iv) distribution, and (v) environmental determinants of the global distribution of clusters.

4.3.1. Ordination of environmental variables

Principle component (PC) analysis of environmental variables revealed that 55.5% of the variation could be explained by the first two axes (**Figure 4.5A**). The most important environmental variables in the delineation of stations in the ordination space were latitude, Lyapunov, fluorescence, flux, angular scattering and phosphate, indicated by the length of the arrow. PC axis 1 differentiates high abundant stations from low abundant stations, accounting for 36.3% of the variance in the data. The environmental variables delineating stations along PC1 were phosphate, F_{cdom}, temperature and salinity. PC1 can be related to the environment supporting diatoms, as it was found that all the low-abundant stations were towards the right (increasing value of temperature and salinity while increasing value of phosphate and F_{cdom}, for instance) whereas highly abundant stations were located on the left (decreasing values of temperature and salinity, for instance). This was in agreement with many diatoms being known to prefer environments that are cold, nutrient rich, and with low salinity. PC axis 2 explains 17.2% of the variance in the data, representing a gradient described by increasing NPP and decreasing fluorescence and Lyapunov exponent (**Figure 4.5A**).

Stations from the Southern Ocean (SO) formed a distinct non overlapping cluster, whereas the rest of

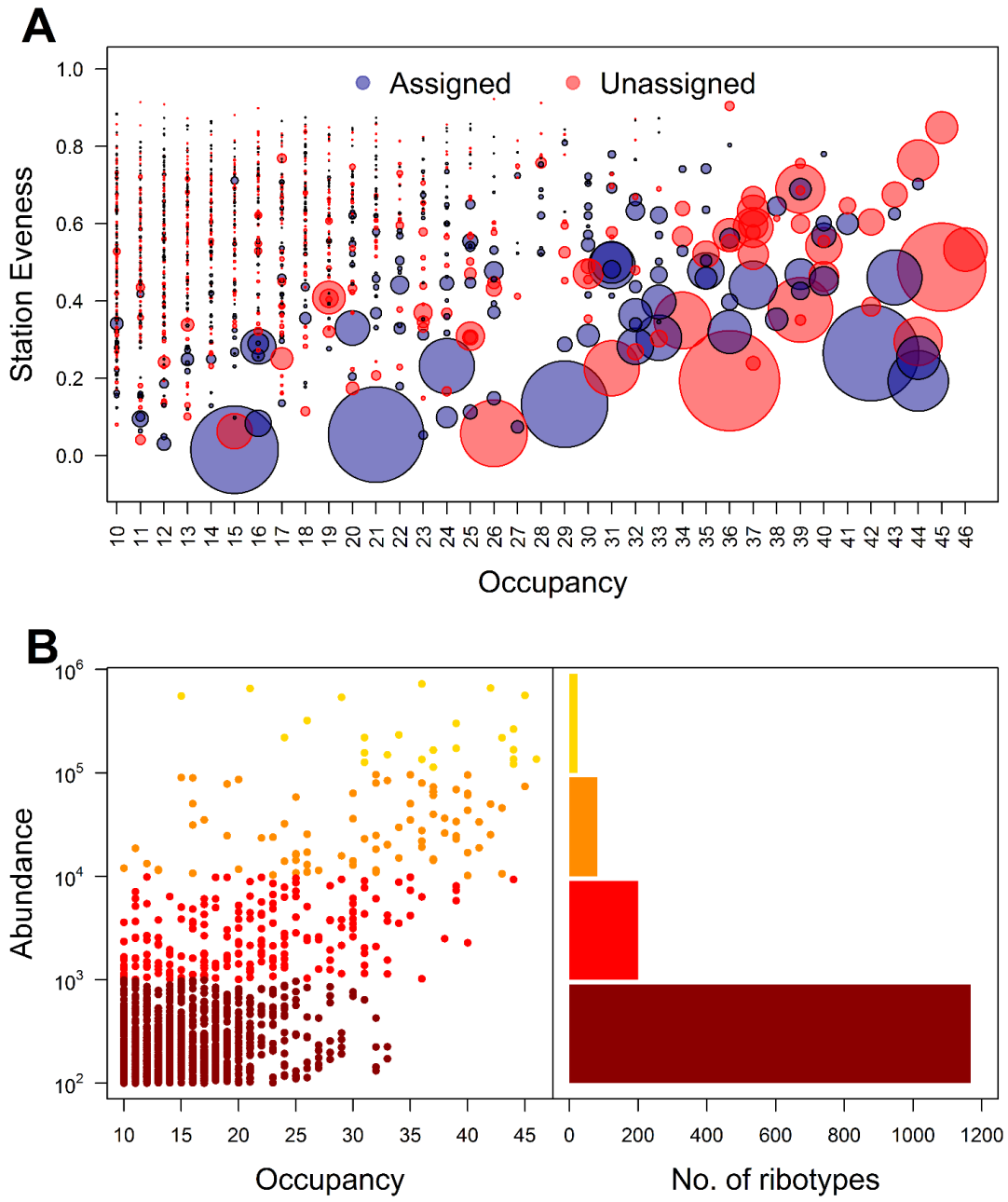


Figure 4.3. Description of dataset used in the study. (A) The plot shows occupancy, cosmopolitanism and station evenness of each ribotype. Each dot corresponds to a ribotype, the larger it is, the more reads it contains. X-axis corresponds to the number of stations in which a ribotype occurs; Y-axis corresponds to the evenness of the ribotype in those stations in which it occurs. 1479 ribotypes were selected based on occupancy (seen in at least 10 stations) and pooled abundance across all stations (≥100 reads). These were designated as 'common ribotypes', representing ~90% of the total reads assigned to diatoms. 46.8% (of 'common ribotypes') could not be assigned upto the genus level using PR2 reference database (for details on taxonomic assignment, please see Chapter 5). (B) Occupancy Vs abundance plot. The color represents the four abundance classes, (C) Number of ribotypes for four abundance classes.

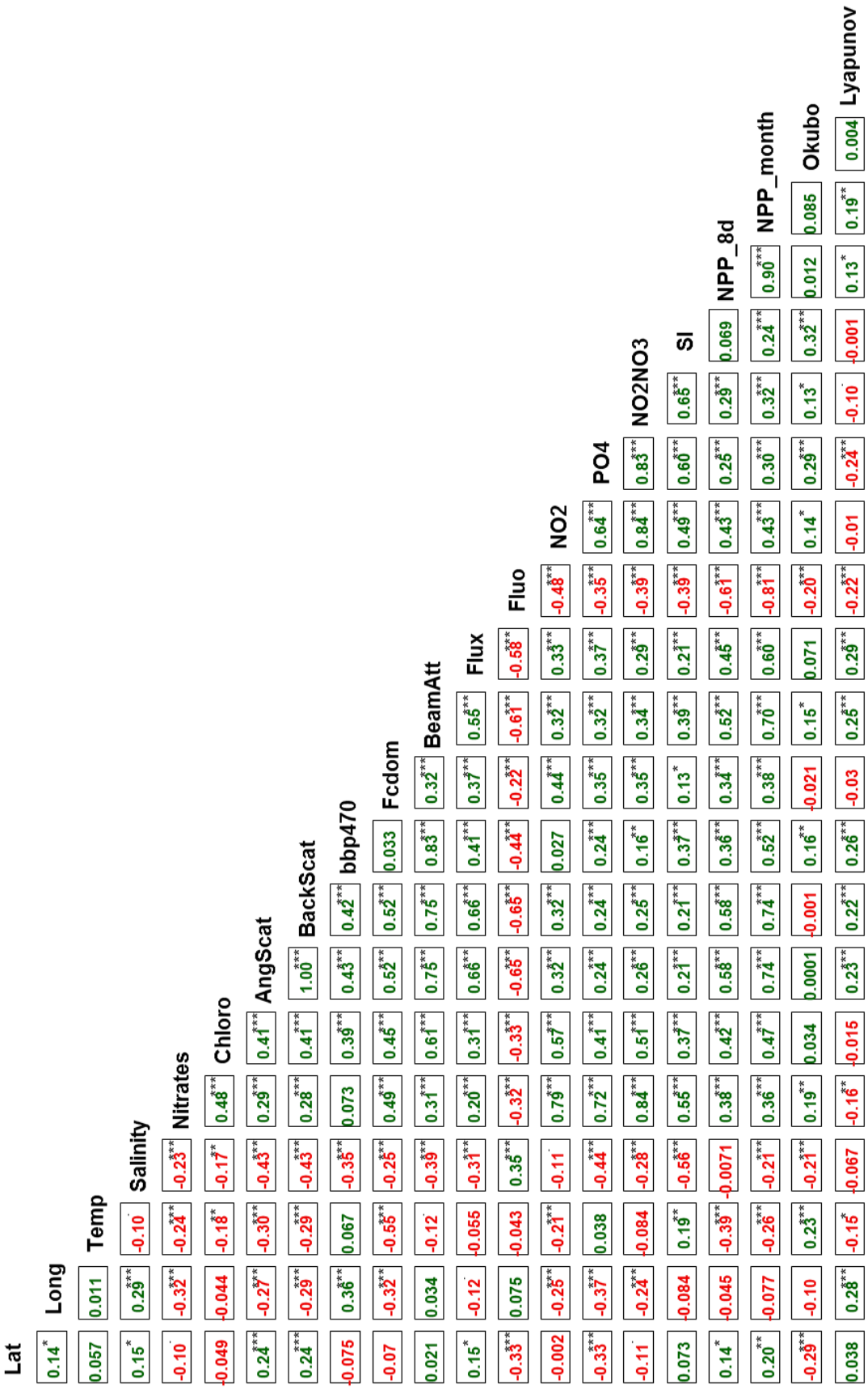


Figure 4.4. (A) Description of environmental variables. All variables were log-transformed to obtain normality. Scatterplot matrix showing Spearman correlation between variables with significance asterisks (0.1[·], 0.05[*], 0.01[**], 0.001[***], 0[*****]). Values in red and green indicates negative and positive correlations respectively.

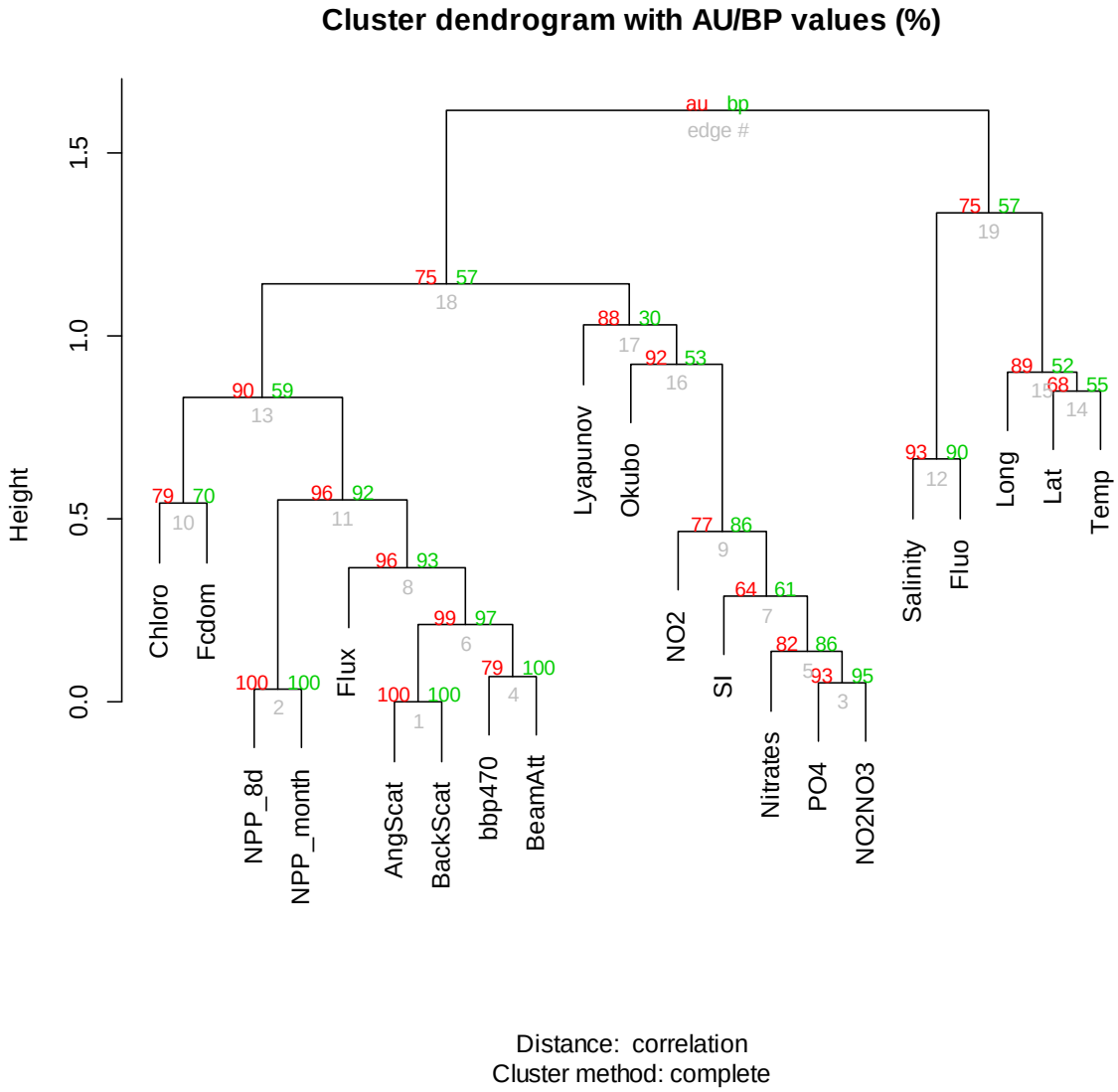


Figure 4.4. (B) Clustering of environmental variables. Lat: Latitude; Long: Longitude; Temp: temperature; Sal: salinity, Chloro: chlorophyll; AngScat: angular scattering; BackScat: back scattering; bbp470: particulate backscattering coefficient at 470 nm; FCDOM: Fluorescent chromophoric dissolved organic matter; Beam Att: Beam Attenuation; Flux: carbon flux; Fluo: fluorescence; NO2: nitrite; PO4: phosphate; NO2NO3: nitrate and nitrite (Inorganic nitrogen); SI: silicate ; NPP_8d: net primary production (8 days); NPP_month: Net primary production (monthly); Okubo: Okubo-Weiss Parameter; Lyapunov: Lyapunov exponent.

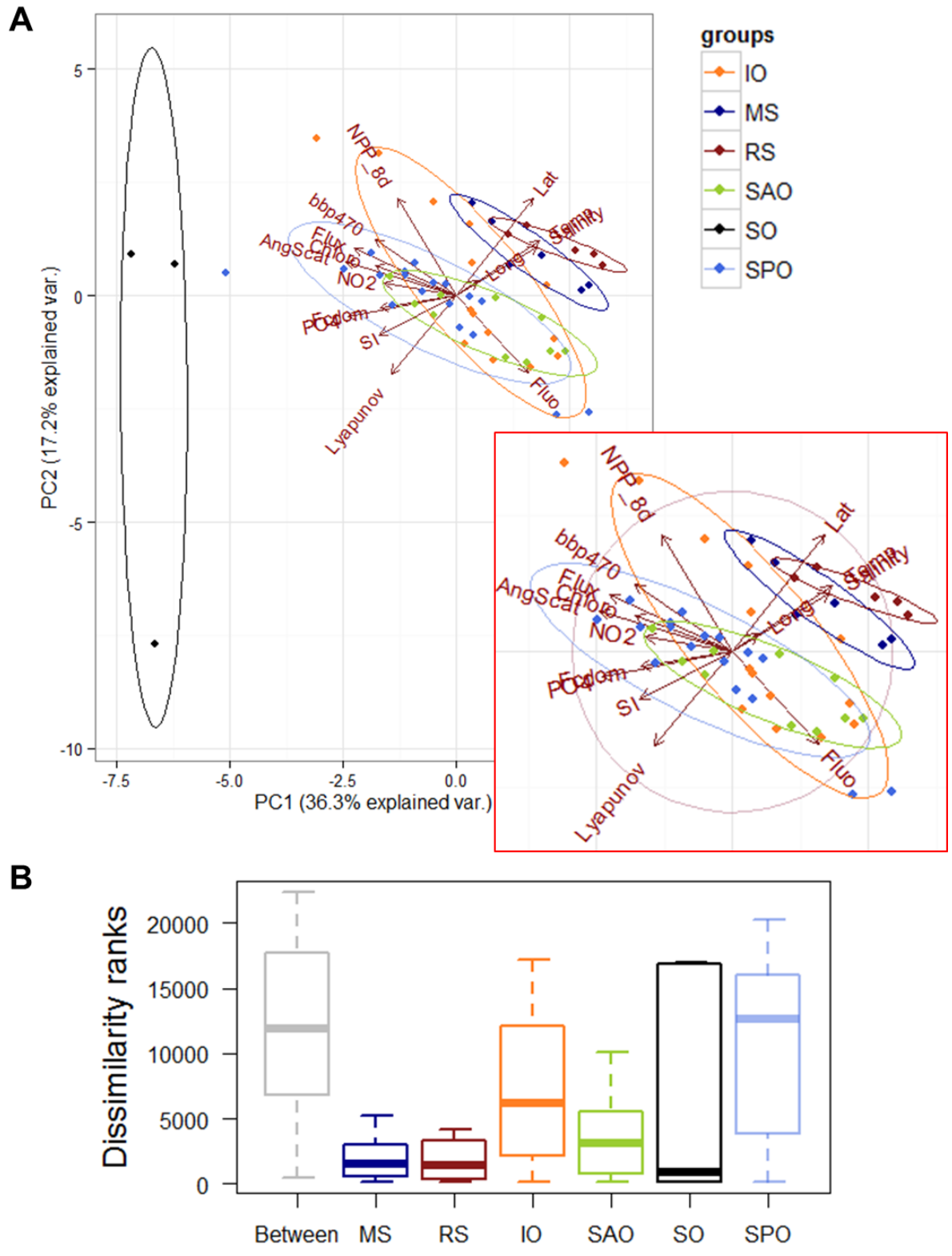
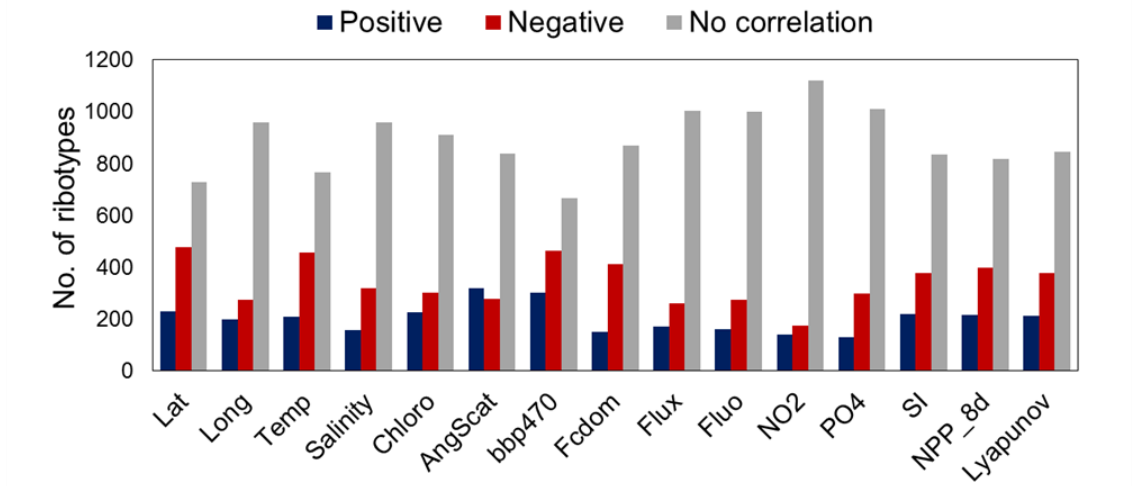


Figure 4.5. (A) PCA ordination of sampling stations. Arrows (eigenvectors > 0.25) indicate the direction of maximum change for the environmental variables. **(B) Analysis of similarities (ANOSIM).** Boxplot showing dissimilarly rank for the one-way ANOSIM between oceanic provinces (999 permutations significance level (p) = 0.001). *MS: Mediterranean Sea; RS: Red Sea; IO: Indian Ocean; SAO: Antarctic Ocean; SO: Southern Ocean; SPO: South Pacific Ocean*.



	Lat	Long	Temp	Salinity	Chloro	AngScat	bbp470	Fcdom	Flux	Fluo	NO2	PO4	SI	NPP_8d	Lyapunov
Positive	230	199	210	158	225	319	302	152	170	160	139	129	221	217	211
Negative	478	276	458	318	301	279	465	413	262	276	176	297	378	399	379
No correlation	727	960	767	959	909	837	668	870	1003	999	1120	1009	836	819	845

Figure 4.6. Multiple regression analysis. Spearman correlations were calculated individually between each ribotype distribution and abiotic variables. [Correlations significant at $p < 0.05$]. The highest number of ribotypes were found to be significantly correlated to temperature. However, silicate and flux showed highest number of ribotypes with positive correlation.

the oceanic provinces exhibited overlap among them. SO was found to be negatively correlated to temperature, latitude and salinity, and positively correlated to phosphate, nitrate, F_{cdom}, silicate, flux and chlorophyll. SAO and SPO exhibited major overlaps, indicating similar environmental heterogeneity. Also MS and RS were quite similar; the stations sampled in these provinces being characterized by high latitude, high temperature and high salinity. Stations sampled in IO could be divided into two distinct groups along the PC2. The first was found to be highly similar to SAO and SPO, whereas the second was very different from others and was structured along the increasing gradient of NPP.

The analysis of similarities (ANOSIM) affirms that the differences between oceanic provinces as illustrated by PC analysis are significant ($R = 0.649$, $p = 0.001$). Based on the (rank) similarity matrix underlying the ordination of samples (Clarke and Warwick 2001), this non-parametric permutation procedure (Monte Carlo tests, Hope 1968) accounts for the variability between stations grouped *a priori* to the respective oceanic provinces. The one-way ANOSIM test (**Figure 4.5B**) shows that different stations within one province are more similar to each other than to any stations from different provinces, although substantial overlap was seen. The Indian Ocean diatoms occupy an intermediate position with a large difference between northern and southern Indian Ocean stations. The remaining combinations within the Indian Ocean show that their environmental characteristics overlap. The Southern Atlantic Ocean also constituted a heterogeneous groups of stations, with Stations 66 and 68 sharing a substantial band of common features which were distinctly separated from the other SAO stations.

4.3.2. Correlation of individual variables to each ribotype

The correlation computed between each individual ribotype and environmental variable indicated that significant relationship exist only for few ribotypes. For temperature, only 46.6% of the ribotypes showed a significant relationship and the majority were negatively correlated. On the other hand, many significantly related ribotypes exhibited negative correlation with latitude, bbp470, F_{cdom}, silicate, NPP and Lyapunov. About 14.8% of the ribotypes displayed no significant correlation to any variable (**Figure 4.6**).

4.3.3. Taxonomic and environmental characterization

Ward's hierarchical clustering of common diatom ribotypes based on Pearson's correlations defined nine clusters (**Figure 4.7**). Kendall's coefficient of concordance affirmed a fair degree of significant agreement among the ribotypes of each cluster. All nine clusters could be broadly divided into three

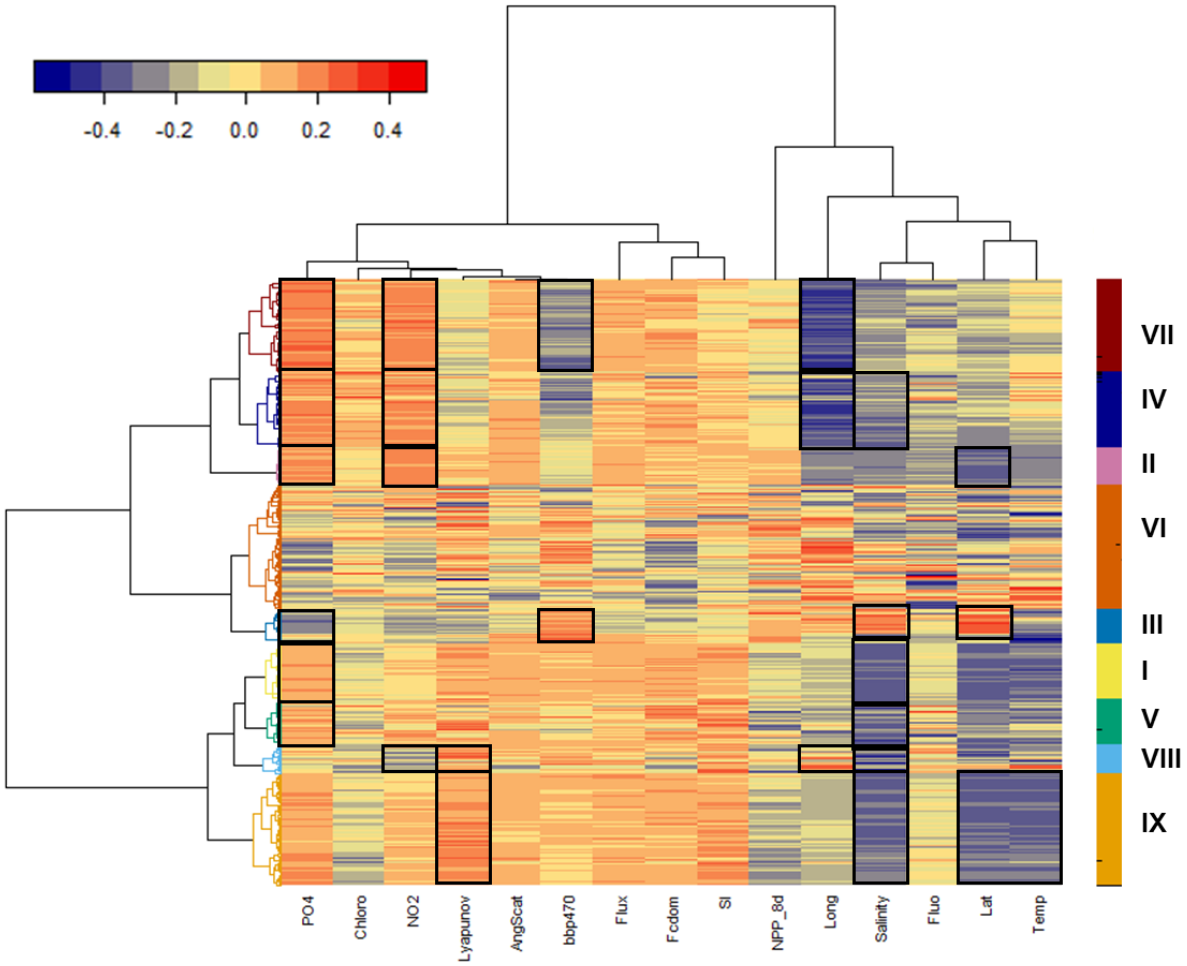


Figure 4.7. Clustering of "common ribotypes" and their correlation to environmental variables. Significant Spearman's correlations are depicted between the ten variables and selected "common ribotypes" (red-negative, green-positive). Nine major clusters were identified, namely, I(132), II(87), III(80), IV(178), V(106) , VI(296), VII(220), VIII(71) and IX(265).

categories as follows:

Category A - **Phosphate$^{(+)}$nitrate$^{(+)}$longitude$^{(-)}$ assemblages**: e.g., Clusters II, IV, VII.

Category B - **Salinity$^{(-)}$temperature$^{(-)}$latitude$^{(-)}$Lyapunov$^{(+)}$ assemblages**: e.g., Clusters I, V, VIII, IX.

Category C - **Variable assemblages**: Those with no common trend. e.g., Clusters III, VI.

The number of ribotypes and certain common characteristics identified for each cluster are described below (**Figure 4.7, Figure 4.8 and Table G4.2**):

Cluster I is composed of 132 ribotypes and shows a significant negative correlation with salinity and latitude and a weak positive correlation with phosphate. It was characterized by the presence of *Chaetoceros, Thalassiosira tumida, Thalassioisira weissflogii* and *Proboscia alata*, as well as unassigned ribotypes.

Cluster II is composed of 87 ribotypes and is correlated negatively with latitude and positively with phosphate and nitrate. It represents a strict polar centric cluster of uncertain identity.

Cluster III is composed of 80 ribotypes and correlates positively with salinity, latitude and bbp470. It was characterized by *Chaetoceros rostratus, Proboscia alata,* and *Leptocylindrus.*

Cluster IV is composed of 178 ribotypes and shows a strong negative correlation with longitude and positive correlations with phosphate and nitrate. It was characterized by *Actinocyclus curvatulus* and *Pseudo-nitzschia* and a large number of unassigned raphid pennates.

Cluster V is composed of 106 ribotypes and shows a significant negative correlation with salinity and a positive correlation with phosphate. It was characterized by *Actinocyclus curvatulus*, *Chaetoceros rostratus* and *Corethron.*

Cluster VI is the largest cluster and is composed of 296 ribotypes. This cluster does not show any common characteristic as every member ribotype reacts differently to the broad scale of environmental gradients. It was characterized by *Actinocyclus curvatulus* and *Thalassiosira*, as well as many unassigned polar centrics and raphid pennates.

Cluster VII is composed of 220 ribotypes and shows a significant negative correlation with longitude and positive correlations with phosphate and nitrate. It was found rich in *Thalassioisira punctigera* and unassigned *Thalassiosira*, together with many ribotypes that could not be assigned.

Cluster VIII is the smallest cluster and is composed of 71 ribotypes showing a weak but significant negative correlation with nitrate. It was characterized by *Corethrom inerme*.

Cluster IX is composed of 265 ribotypes and shows a significant negative correlation with salinity, temperature and latitude, and positive correlations with Lyapunov exponent. It was characterized by *Fragillariopsis* and *Chaetoceros.*

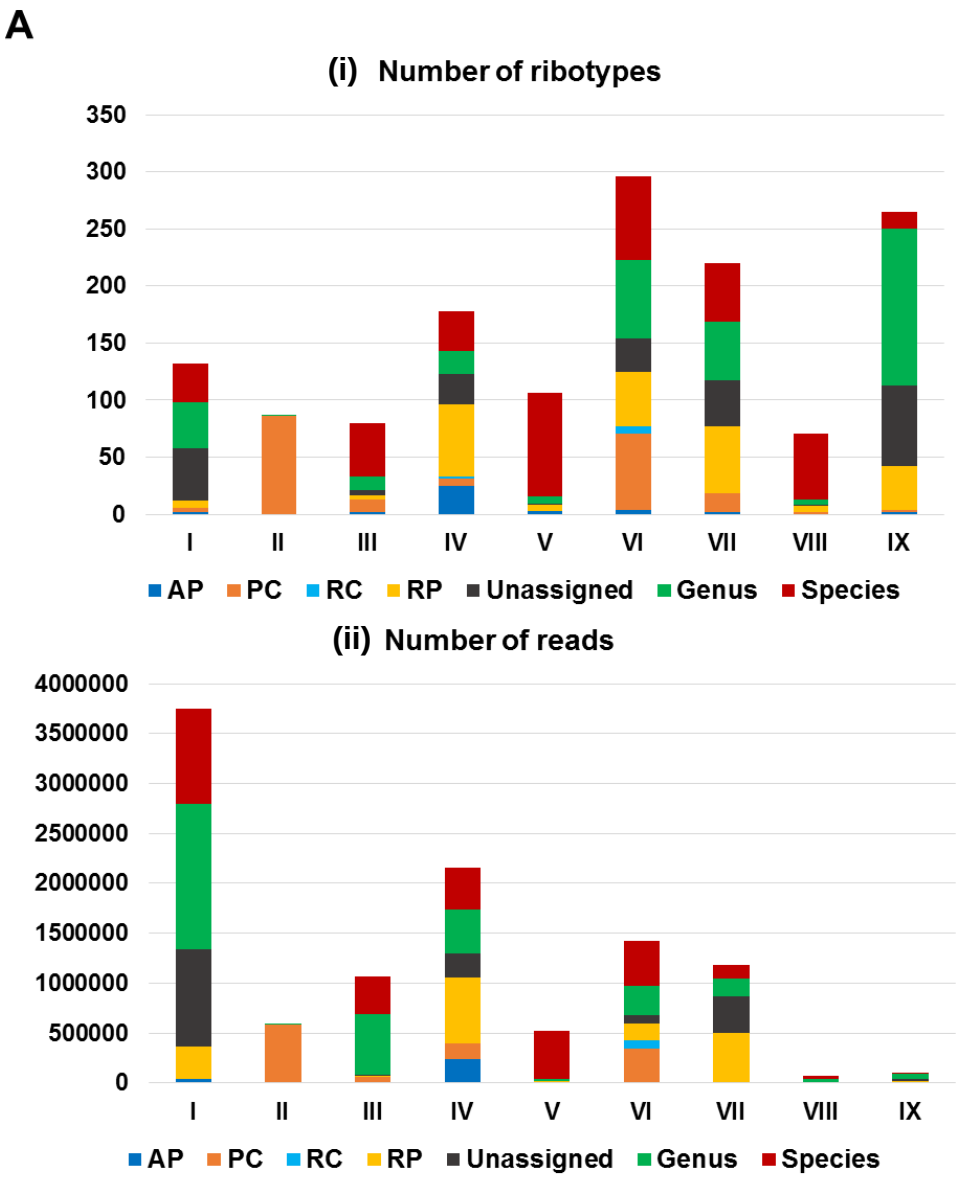


Figure 4.8A. How novel is a cluster? Taxonomic Composition of each cluster based on (i) richness (unique ribotypes) and (ii) abundance (total reads). I-IX represents nine clusters identified in the study (for detail, please refer Fig 4.7). Bar color corresponds to the color of clusters reported in **Fig 4.7**.

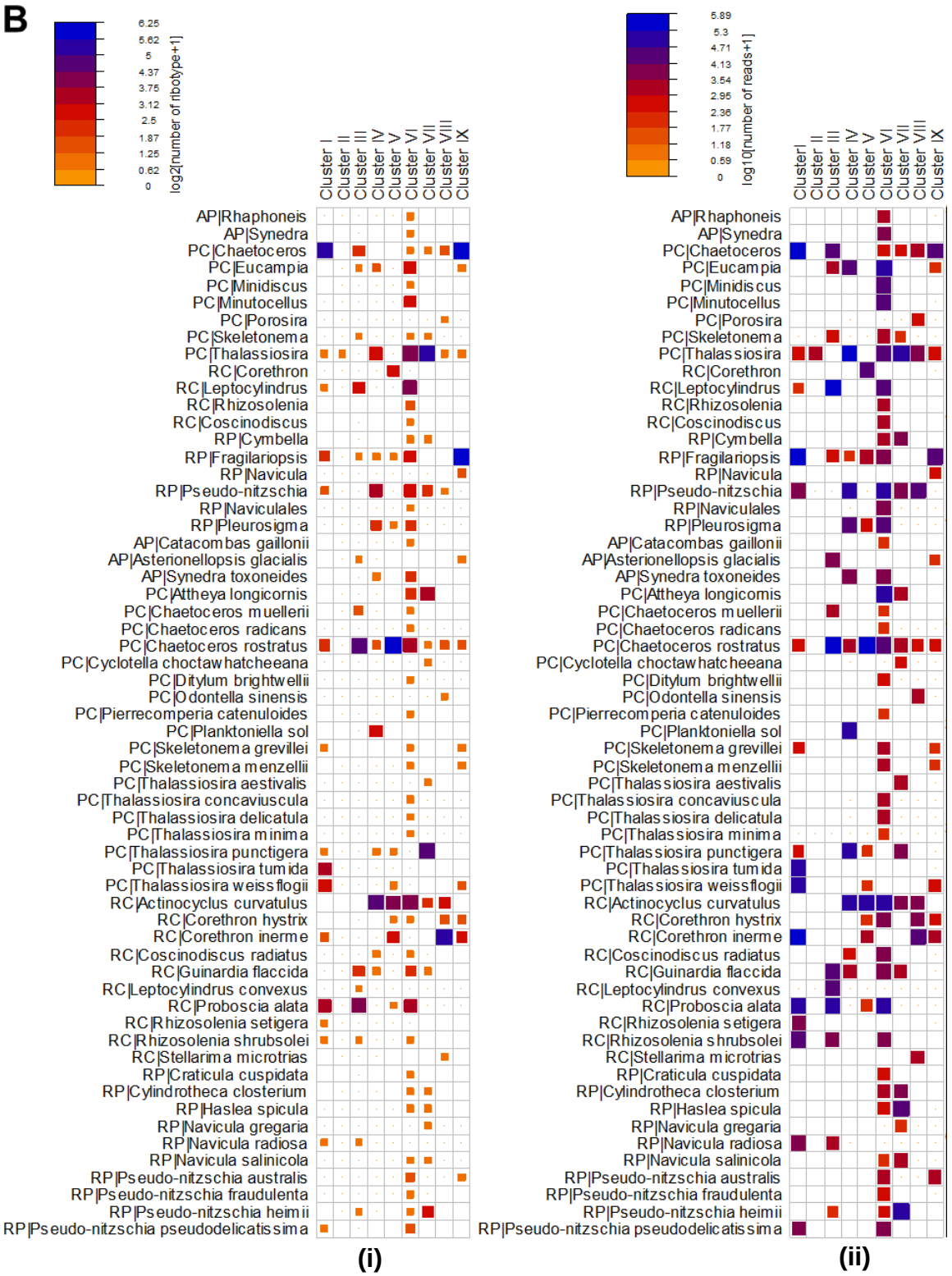


Figure 4.8B. Characteristic genera and species of each cluster based on assigned fraction. (i) Richness. (ii) Abundance.

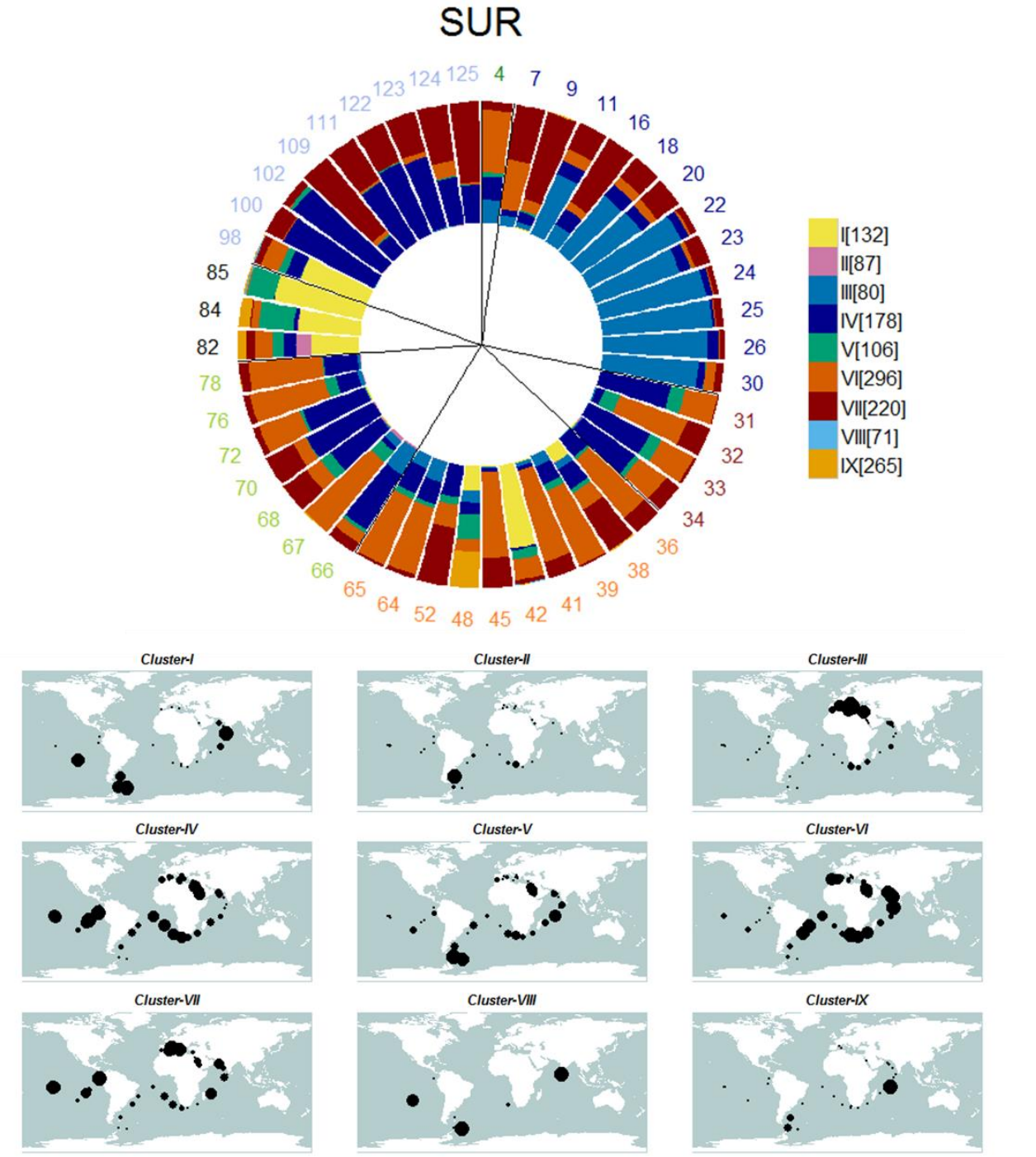


Figure 4.9A. Distribution of clusters at surface. ***Upper Panel*****:** Proportional distribution of pooled abundances of clusters across stations. Key for color-coded clusters is shown. Numbers in the parentheses indicate the number of ribotype members in each cluster. ***Lower Panel*****: Map showing the distribution of reads across stations for each cluster.** The size of the bubble is proportional to the number of reads it contain, and is not the same for each cluster.

Focusing on the unassigned to assigned fraction of members, it was found that the ratio of unassigned to assigned within clusters I, IV, VI, VII and IX were nearly 1:1. On the other hand, Cluster II was found to be a strict polar centric cluster whereas Clusters III, V and VIII displayed a majority of assigned ribotypes. Unassigned polar centrics were seen principally in Clusters II and VI. Unassigned raphid pennates were clustered under IV, VI and VII. Altogether unassigned araphid pennates represent only 3% of the common ribotypes and were seen principally in Cluster IV (**Figure 4.8A and Table G4.2**).

4.3.4. Spatial characterization at local scale

Each surface station exhibited a distinct composition (reads pooled over clusters for surface samples). The stations were organized under nine defined clusters based on their major component, as indicated below (*upper panel* in **Figure 4.9A**):

Cluster I : Stations 42, 82, 84, 85, 98.

Cluster II : Not a major component, except at Station 82.

Cluster III : Stations 11, 18, 20, 22, 23, 24, 25, 26, 30.

Cluster IV : Stations 31, 33, 34, 66, 68, 70, 72, 100, 102, 111, 122, 123.

Cluster V : Not a major component.

Cluster VI : Stations 4, 32, 36, 39, 41, 45, 64, 65, 67, 76, 78.

Cluster VII : Stations 7, 9, 16, 38, 52, 109, 124, 125.

Cluster VIII : Not a major component, although found predominantly at Stations 42, 85 and 98

Cluster IX : Station 48.

Figure 4.10 shows the location of each station, color-coded based on the dominant cluster in the surface sample of each station. Based on the above outlined geographic preference common to most of the stations in a province, a generalized spatial depiction can be suggested for each cluster as shown in **Figure 4.9A** (*lower panel*). All the clusters exhibited almost similar geographical preferences at both depths (i.e. SRF and DCM) **Figure 4.9B.**

4.3.5. Spatial characterization at regional scale

The total number of reads within each cluster indicated that three clusters (II, VIII and IX) are represented by only a few reads whereas two of them (I and IV) were many times (2-10 times) greater than the rest (*upper panel* in **Figures 4.11**). Despite this, all the clusters were found to be remarkably dominant in SPO, particularly IV and VII. Clusters I and V related well with SO. Cluster III and VI were primarily seen in SPO but also exhibited signals for MS and IO, respectively (*lower panel* in **Figure 4.11**). Based on their relative distribution across provinces, three classes can be defined, namely SO-SPO

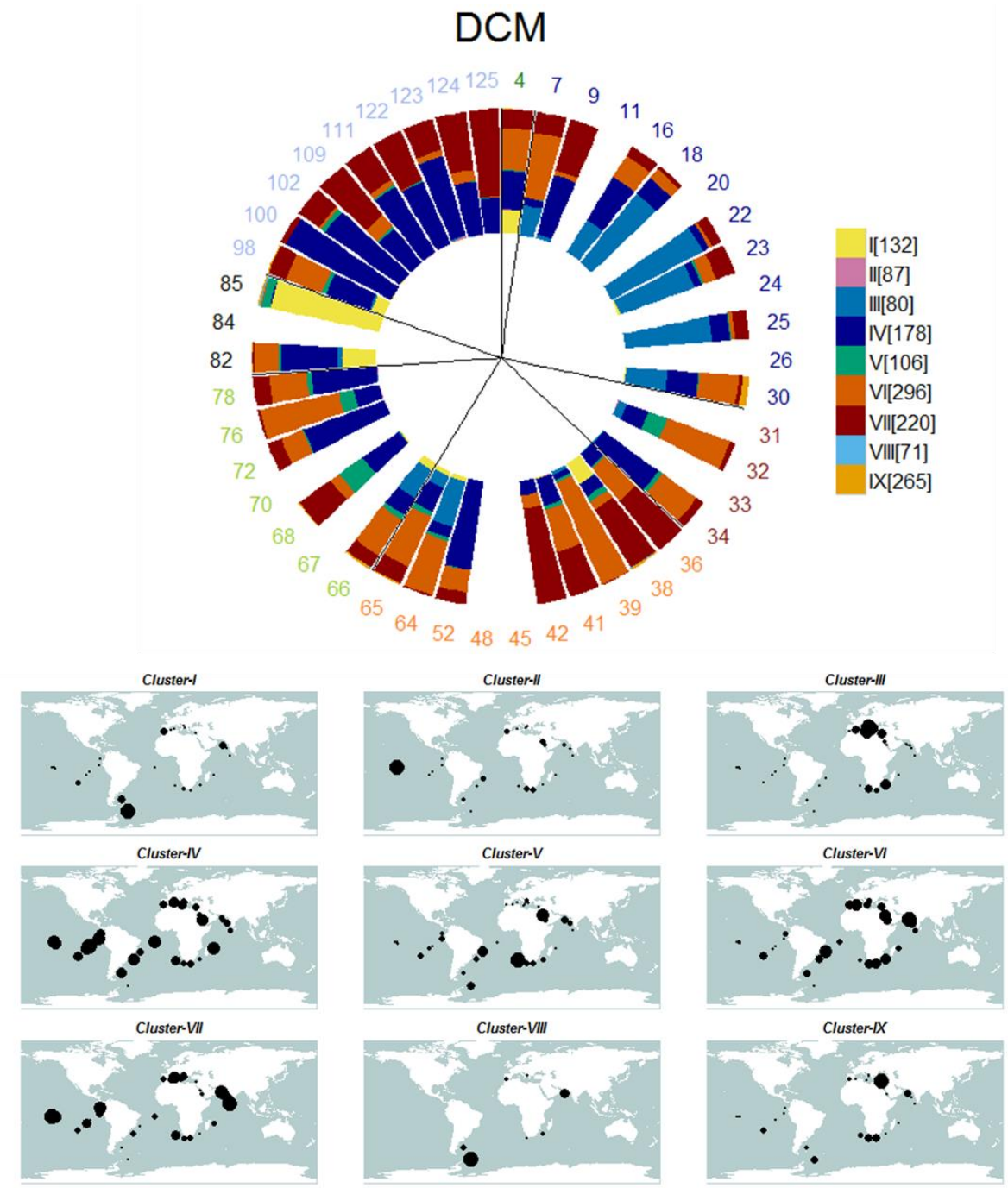


Figure 4.9B. Distribution of clusters at DCM. ***Upper Panel*****:** Proportional distribution of pooled abundances of clusters across stations. Key for color-coded clusters is shown. Numbers in the parentheses indicate the number of members in each cluster. ***Lower Panel*****: Map showing the distribution of reads across stations for each cluster.** The size of the bubble is proportional to the number of reads it contains and is not proportional between different clusters.

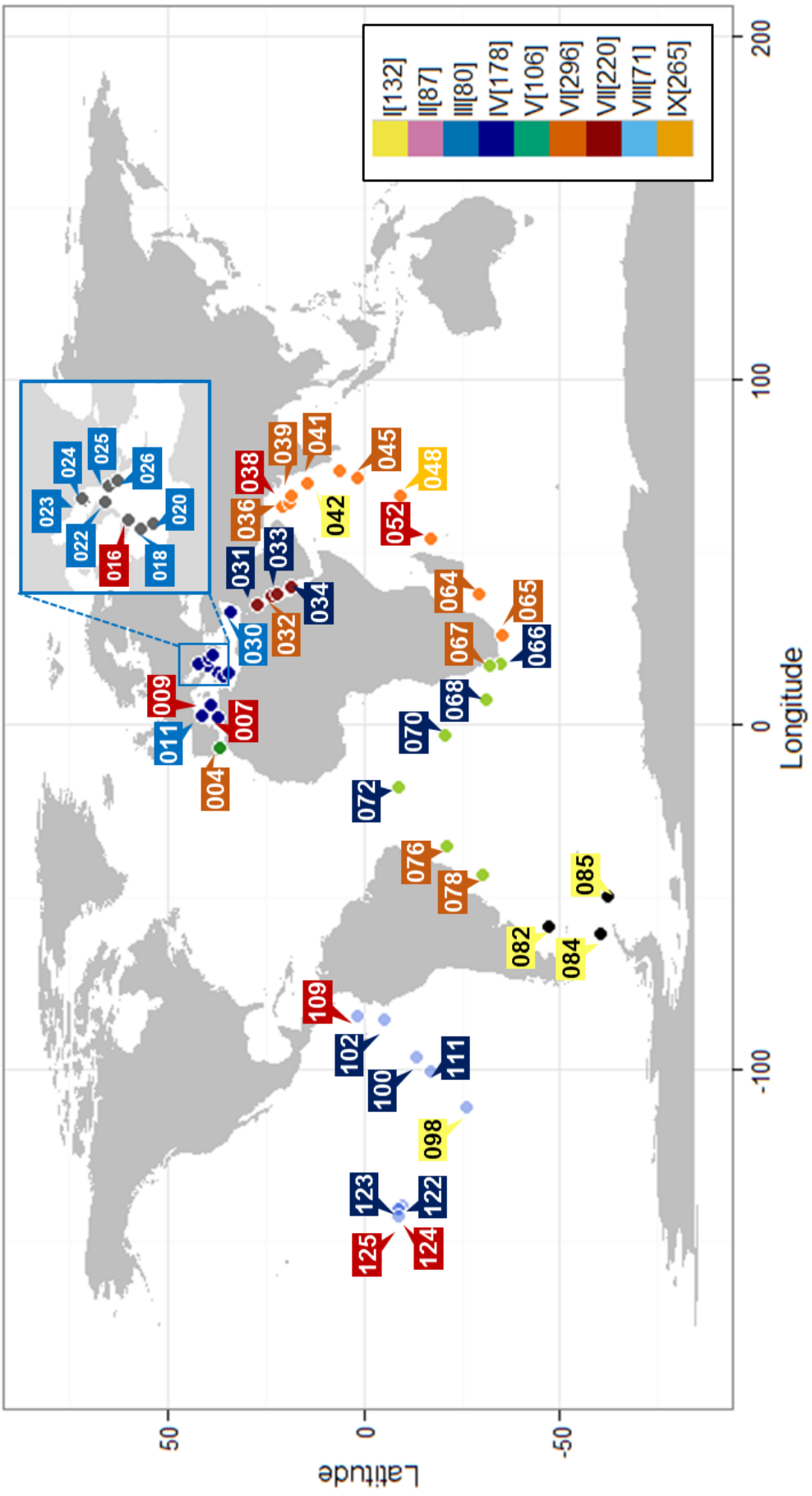


Figure 4.10. Tara map colored by the dominant cluster at the surface waters of each station. Key for color-coded clusters is shown. Numbers in the parentheses indicate the number of ribotype members in each cluster.

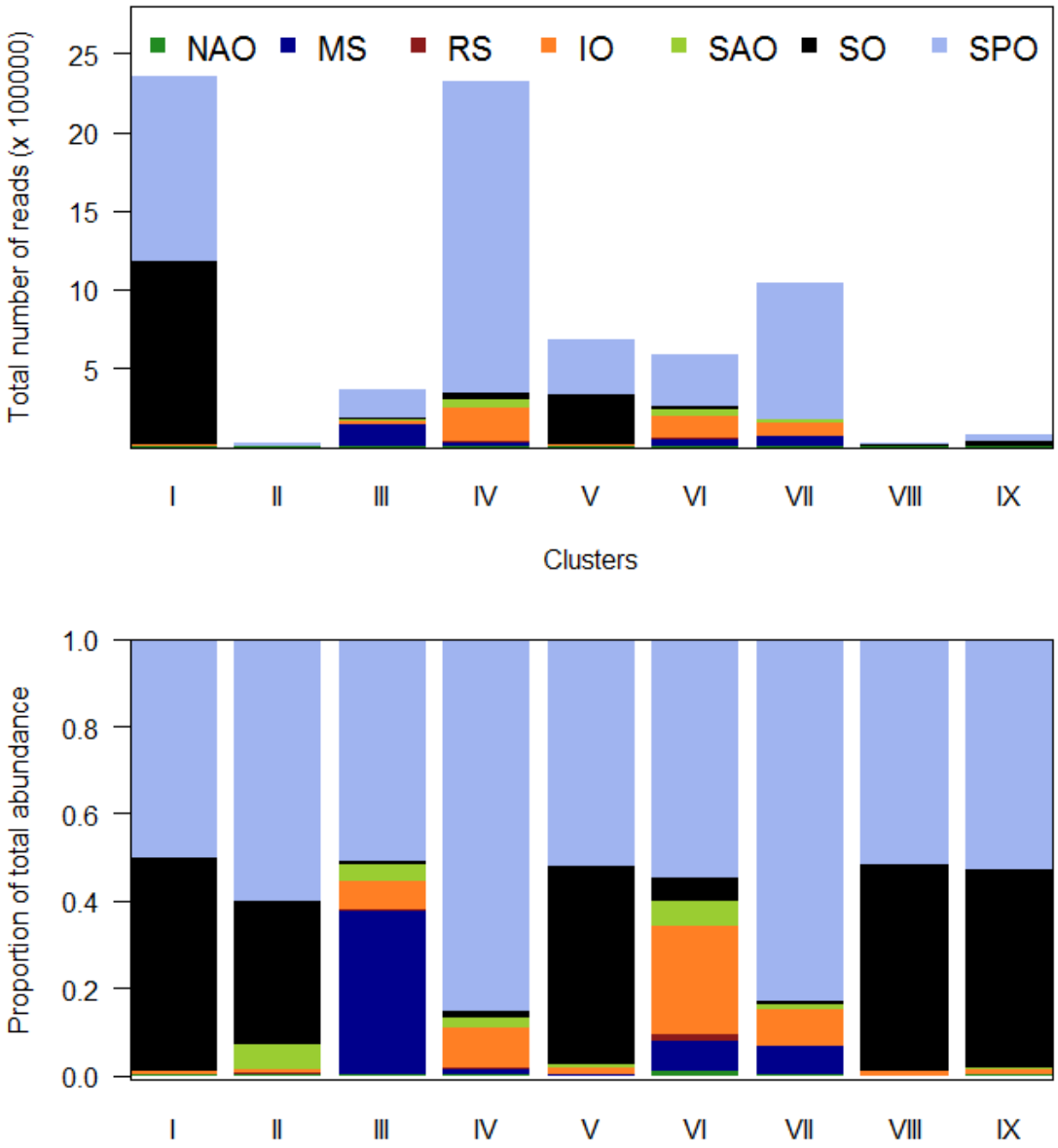


Figure 4.11. Distribution of pooled abundances of clusters in oceanic provinces. I-IX represents nine clusters identified in the study (for detail, please refer Fig 4.7).

clusters (I, II, V, VIII, IX), SPO clusters (IV and VII), and cosmopolitan clusters (III and VI) (*lower panel* in **Figure 4.11**).

4.3.6. Environmental determinants of the global distribution of clusters

After defining clusters (ribotype associations), key environmental drivers were identified for each cluster. The percentage variation explained ranged between 8.31-68.3% (**Figure 4.12**); the lowest being in Cluster VIII and the highest corresponding to Cluster III. In the largest two clusters (I and IV), the total explainable variation was 39.78% and 29.31%, respectively. The main drivers for these clusters were latitude and temperature for Cluster I and longitude and bbp470 for Cluster IV. Cluster III was the one with maximum explained variation (68.3%) and the majority of variation was explained by temperature and latitude. Latitude explained the most variation in Cluster V. Longitude was found to be majorly explaining the variation in Clusters VI and VII with F_{cdom} and Fluorescence respectively. For the three smallest clusters, i.e., II, VIII, IX, the key drivers were latitude, salinity, and, latitude and Lyapunov, respectively (**Figure 4.12**). For each cluster, the variation explained by environmental variable, expressed as percentage contribution, is reported in **Table G4.3**.

The impact of each individual environmental variable in best describing a cluster was assessed by scaling each variable by the maximum explained variation of that variable. For instance, for latitude, the percentage explained for each cluster was divided by 13.06 (see **Table G4.3**). The radar-plot presented in **Figure 4.13** suggests that latitude had the major impact on III and V, and the least on Clusters IV and VIII. Similarly, temperature and salinity primarily controlled Cluster III. Angular scattering and bbp470 were seen to control Cluster I and, Clusters III and IV, respectively. Flux was found to be the most influential factor for Clusters I and III. NPP, NO2, PO4 and Lyapunov were comparatively important in Clusters I, IV, III and IX. Silicate had the strongest impact on Clusters III, VI and IX (**Figure 4.13**).

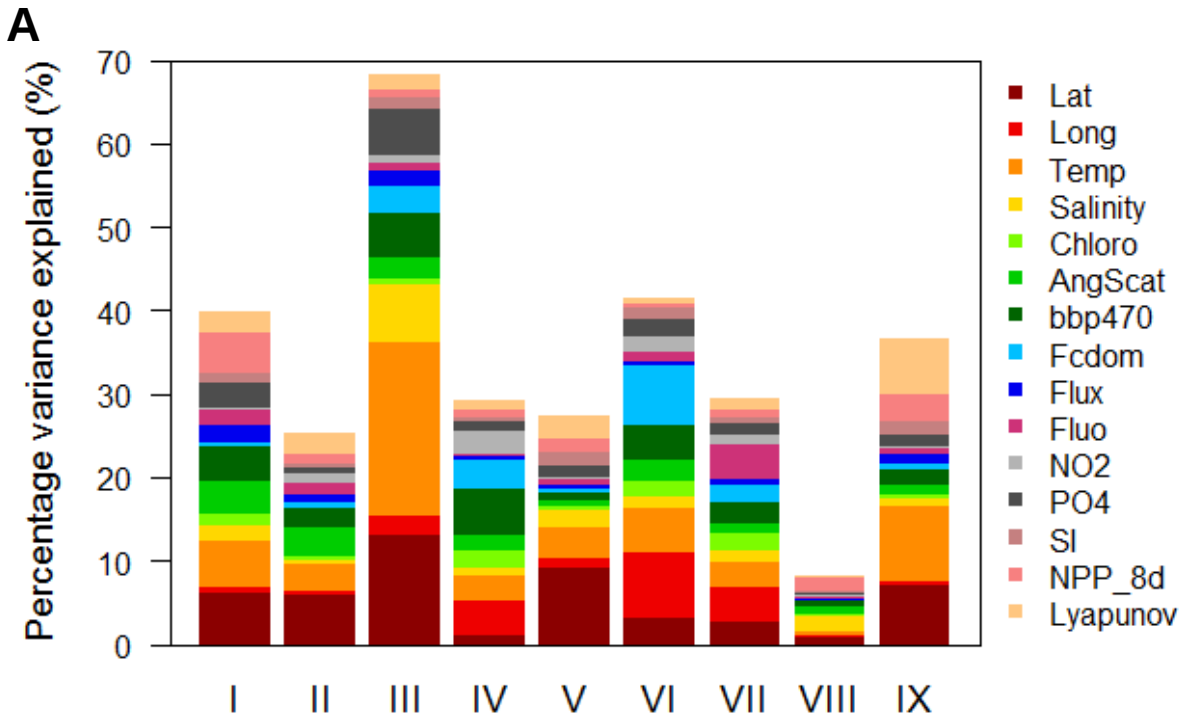


Figure 4.12. Quantifying the contribution of selected environmental variable to a multiple linear regression model. (A) The bar chart depicts the percentage variation explained by each variable in each case. The height of the bars represent the total explained variation. ***Lower panel:*** Radar plot depicting the importance of each variable in a cluster.

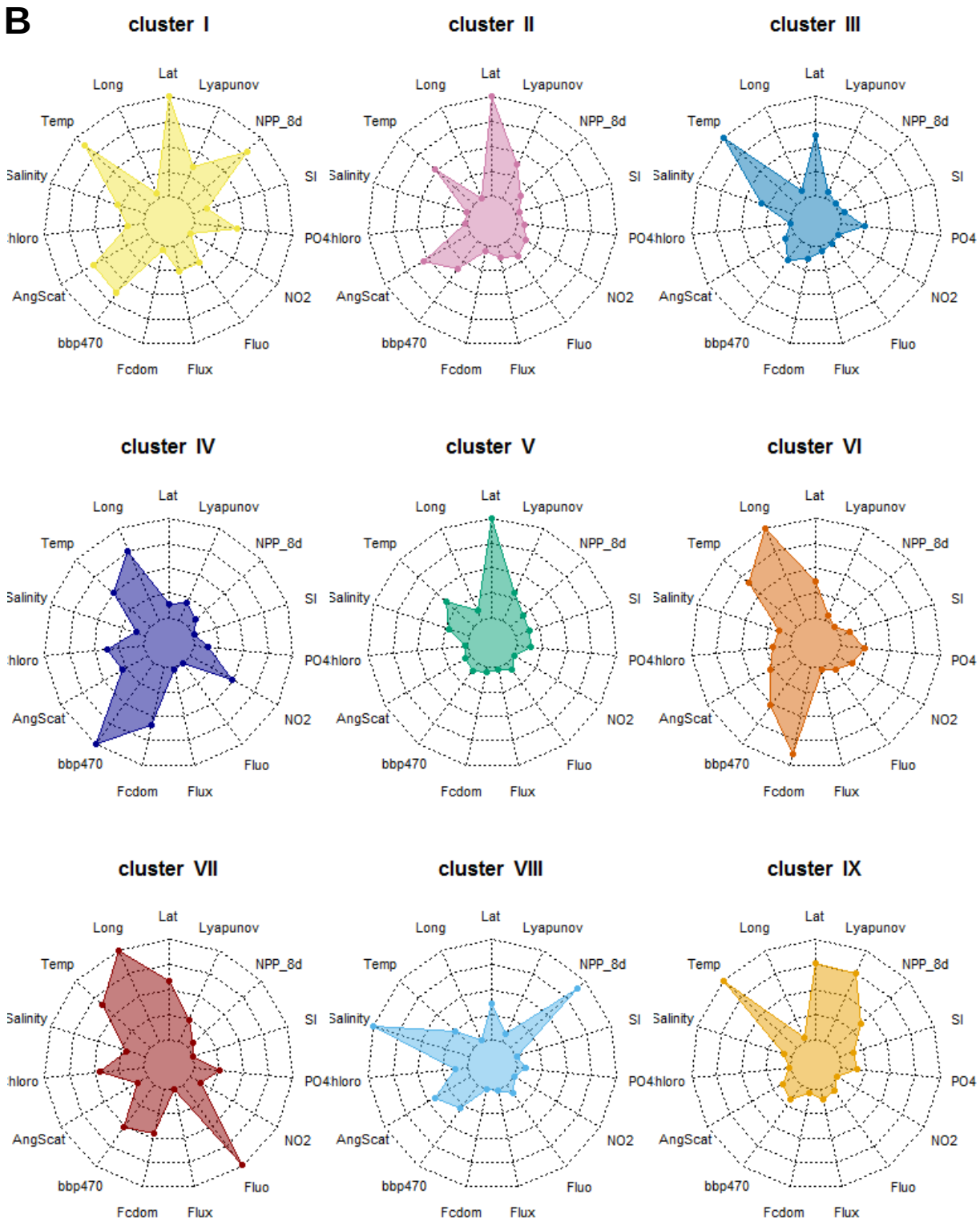


Figure 4.12. Quantifying the contribution of selected environmental variable to a multiple linear regression model. (B) Radar plot depicting the importance of each variable in a cluster.

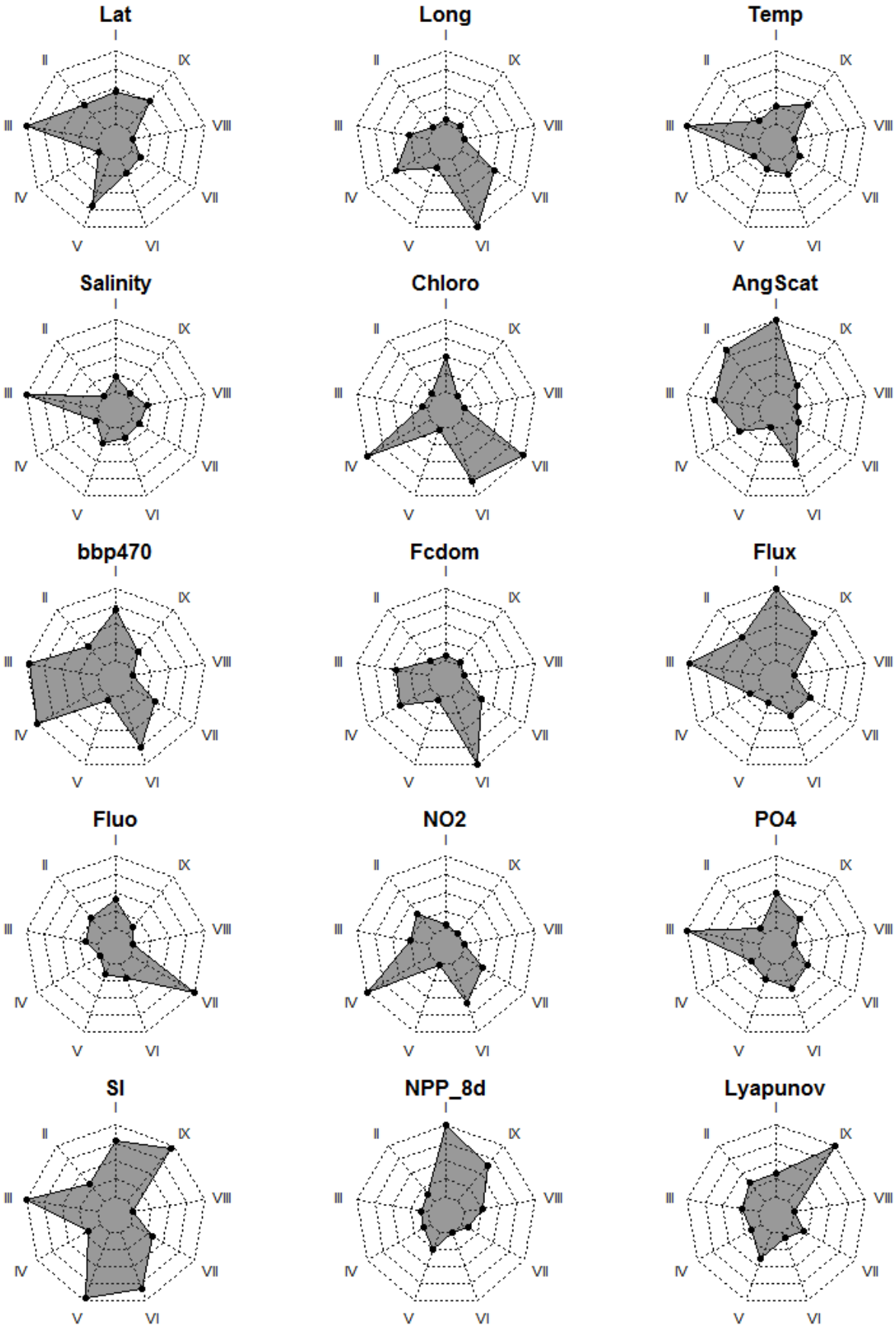


Figure 4.13. Relative importance of each environmental variable across nine clusters. Radar plot depicting the relative variation explained by each variable across all clusters (w.r.t. maximum variation explained by a respective variable among all clusters).

4.4. Discussion

Despite the large number of studies carried out in the recent past on the effect of human pressure on diatom communities (Pan et al., 2000; Soininen, 2002; Potapova and Charles, 2003), the number of studies attempting to characterize the natural patterns and the relative weights of environmental parameters influencing this natural variability is still limited (Sabater and Roca, 1992; Stevenson, 1997). Various independent studies have reported the association of an array of environmental factors with the distribution of different diatom species. It has long been known that temperature, salinity and nutrients have a significant impact on the local abundance pattern of diatoms (e.g., Butcher, 1947; Fjerdingstad, 1950; Zelinka & Marvan, 1961). However, at a global scale these abiotic parameters fail to explain the variation present in community structure. This study was performed with the aim to identify a common set of properties (spatial/environmental) for a co-occurring group of species. The analysis demonstrates that most ribotypes were insignificantly correlated to different variables. However, for those which were significantly correlated, temperature, bbp470, and latitude were the most important ones in controlling their distribution (**Figure 4.6**).

In the present study, cluster analysis was used to define clusters. The total number of reads for each cluster changed remarkably between stations, particularly from different provinces, indicating varied geographical preferences and/or ecological requirements. An evaluation of the influence of environmental variables on each of them suggested that although environmental heterogeneity influences their distribution, it could not completely explain the differences between clusters. This may be due to the point that all factors interact in a definitive manner to give a pattern rather than by acting at an individualistic level. Another explanation may be that these patterns can occur as a direct consequence of biotic interactions such as mutualism, competition and predation (Roxburgh and Chesson, 1998).

Nine clusters displayed a varying number of ribotypes ranging from 71-296 ribotypes and their dominance was not related to their richness. There was some overlap in genus/species among different clusters, for example, the occurrence of *Actinocyclus curvatulus* among the dominant species of Clusters IV, V and VI. It is noteworthy that a few ribotypes for a particular species does not always mean a low number of reads for that species. For instance, in Cluster I, there were only very few ribotypes assigned as *Corethron inerme*, but the total number of reads represented by those few ribotypes was extraordinarily high.

In each cluster, the numerically dominant species belonged to the most abundant diatoms known, like

Chatoceros, *Thalassiosira, Fragilariopsis* and *Corethron*. *Chaetoceros* was represented by nearly 13% ribotypes totaling to 14% reads. There were three species of *Chaetoceros*, e.g., *C.muellerii, C.radicans, C.rostratus*; along with unassigned *Chaetoceros*. All the clusters seem to have either of the variants, although Clusters I, III, V and IX were dominated by either unassigned *Chaetoceros* (I and IX) or *C. rostratus* (III and V). The other two species were represented by only a few ribotypes (and reads). *Thalassiosira*, *Fragilariopsis* and *Corethron* were the other genera which were abundant in the dataset. These along with *Chaetoceros* constituted the most abundant cluster, i.e. Cluster I. The environmental variables also indicted this to be a Southern Ocean cluster; negatively correlated to latitude and temperature. The second smallest of the nine clusters (Cluster III) was considered most typical of Mediterranean waters. This cluster contained a majority of reads from *Leptocylindrus, Chaetoceros rostratus and Proboscia alata*. Other diatom clusters also demonstrated regional differences and a dominant role of mostly a single resource.

In addition to a very few numerically dominant species, each cluster had many species that were seen in low abundances. In spite of ecological requirements similar to the abundant ones, a systematic evaluation of individual species traits is desirable to understand their low abundance. Several trait-based analyses (e.g., Bremnera et al., 2006) have been proposed in recent past providing a link between species, environments and ecosystem processes. Thus, using a summary of the biological trait composition of clusters, it is possible to gain a valuable explanation of ecological functioning. Integrating a range of traits from those that are closely linked to important ecosystem processes to the ones that are sensitive to anthropogenic impacts will potentially lead to a comprehensive description of ecological functioning.

CHAPTER 5

Discerning and Quantifying Power-law Behavior of Protistan Communities

Summary

Abstract

This study presents an approach to discern and quantify the distribution of rare species in the protist community data sets generated by *Tara* Oceans. The plotted rank abundance distribution (RAD) for all the samples under study showed a long-tailed distribution whose tail appears to follow a power-law behavior. This work explores the patterns of species abundance by employing rank abundance distributions along with commonness and rarity patterns of protists in the world's ocean. In addition, I assessed the ubiquity of a ribotype across sampling sites to understand the extent to which a rare ribotype remains rare in space.

5.1. Introduction

> *"The answers to general ecological questions are rarely universal laws, like those of physics. Instead, the answers are conditional statements such as: for a community of species with properties A1 and A2 in habitat B and latitude C, limiting factors X2 and X5 are likely to predominate."* *-Diamond and Case, 1986*

A major research aim in ecology has been to understand the mechanisms and processes that generate and shape the differences among species abundances and distribution (Whittaker 1965, 1970, 1972; McGill et al. 2007). This is fundamental to understand community structure, biology and ecology, thereby facilitating their characterization. It has been seen as a universal feature that each community is characterized by a few abundant, some moderately common, and many uncommon or rare species. However, different models predict that, when ordered according to their abundance, species within a community obey different distributions.

5.1.1. Marine community structure: evident structuring processes

Ecological communities are the product of both contemporary biotic and abiotic forces as well as historical (phylogenetic) contingencies (Westoby, 2006). For example, biogeography, local adaptive radiation, intra- and inter-specific interactions together with effects imposed on a community by habitat characteristics dictate the community assemblage. None of the processes are mutually exclusive and their relative strength varies at different temporal, geographical and phylogenetic scales.

The various ecological processes central to community assembly, as reported in literature, can be summarized as follows:

(a) *Competition* has long been considered to provide the mechanism that structures communities. It represents two (or more) species competing to utilize a common resource of limiting supply,

(b) *Habitat filtering* takes place through non-random colonization and invasion determined by environmental characteristics,

(c) *Predator-prey interactions* take place between trophic levels.

Terrestrial and marine ecosystems vary considerably in the processes that organize them. Community structure theories that have been developed to date were developed mainly with regard to terrestrial communities. Cloern and Dufford (2005) have proposed principles of phytoplankton community assembly which can be generalized to the total marine community assembly: (a) Fast and selective grazing, a powerful top-down force to shape phytoplankton communities, (b) Turbulent mixing, a physical process that selects species on the basis of their size and form (Margalef, 1978; Cullen et al., 2002), (c) Mixotrophy, which allows some algal species to tap organic nutrient pools and to function at multiple trophic levels (Bird and Kalff 1986; Estep et al. 1986), (d) Species interactions across trophic levels (Carlsson et al.,1995; Teegarden, 1999; Calbet et al., 2002; Fistarol et al. 2003), (e) Variable life histories, alternating vegetative and resting stages (Dale 2001; Smetacek, 1985), (f) Dispersal and immigration, (g) Large-scale climatic periodicity (McGowan et al., 1998; Chiba and Saino, 2002; McQuoid and Nordberg, 2003; Zingone and Wyatt, 2004; Hallegraeff and Bolch, 1992), (h) Resource Partitioning, and (i) Habitat Heterogeneity.

Out of the above mentioned factors, turbulent mixing is the factor that is exclusively important for the organization of marine communities. The marine life around the world is mostly dependent on patterns of ocean circulation. Thus, the analysis of ocean dynamics is an essential element in such studies. Oceans are not uniformly mixed but are structured in layers with distinct properties (Colling, 1991; Knauss, 2005). Turbulent mixing creates currents and brings the exchange between different water masses leading to heat redistribution and nutrient circulation (Colling, 1991; Knauss, 2005). In addition to the fundamental role of vertical upwelling or downwelling movements in the circulation of nutrients, these movements give birth to retention areas. Other phenomena that play important roles in ocean circulation can be explained by “Ekman transport” patterns (Ekman, 1902). Following the momentum exerted by each layer of the ocean to the water beneath, its movement gets slightly deviated producing spiral structures (Ekman spiral). Thus, ocean mixing and coastal upwelling bring new supplies of nutrients up from deeper waters providing food for deep-dwelling species. When optimal light, temperature, and nutrient conditions coincide, plankton population explosions called “blooms” occur (Lazier, 2004). These sustain ocean life as they increase the availability of organic material, and are responsible for the later enrichment of the ocean floor by means of “marine snow” (Miller, 2004). In addition, this vertical movement allows the flow of particles to ocean depths which is a critical link in the global carbon cycle (Mann and Lazier, 2006).

Phytoplankton play several key roles in the ocean and climate systems and their dominance is a property of their adaptation to the local water properties (Litchman et al., 2007; Giovannoni and Stingl, 2005; Bouman, et al. 2006). The characterization of planktonic communities is a fundamental problem as it raises the question of the predictability of ecosystems. d'Ovidio et al. (2010) reported that fluid dynamics and horizontal stirring have primary roles in shaping and maintaining phytoplankton community structure. They demonstrated that "in confluence regions, water masses with different properties are stirred, creating contrasted physicochemical conditions which favor the emergence of complex community distributions" (d'Ovidio et al., 2010). This recent shift in the view towards complex patterns of environmental variation fits in the scheme of "patch dynamics models" (Pickett and White. 1985), which posits that marine planktonic community organization is shaped by both stochastic disturbance and biotic forces.

The observation that species vary in number and abundance prompted the development of species abundance distribution (SAD) models (Mugurran, 2004) to describe the relationship between the two. SAD is the most frequently studied pattern in ecology and represents the frequency distribution of species abundances in an assemblage. In recent years, it has been used to test different hypotheses about the processes that determine the diversity of an ecological assemblage, notably the assumption that abundance of a species reflects its success at competing for limiting resources (Mugurran, 2004). SADs can be visualized using two main approaches, histograms of species having a number of individuals within exponentially increasing bin widths (Preston, 1948), and plots of ranked species abundances from the most to least common, known as rank-abundance distribution (RAD) curves (MacArthur, 1957; Whittaker, 1965). The latter has gained popularity as one of the most informative approaches (Nekola et al., 2008) to visualize SAD and to investigate community ecology hypotheses.

Why ecologists look at RADs? Alternatively to SADs, abundance observations can be represented in terms of Rank-Abundance Distributions (RADs). These plots typically display the abundances (or relative abundances, sometimes called frequencies) of species as a function of their rank in an ordering going from the most common to the rarest. In addition to SADs, they are used as reporters of ecological processes to which the observed community is subjected. The shape of the RAD can be compared to the species distribution model that can best describe the community. Thus, by looking at these shapes, simple hypotheses can be made about the way species interact within communities, and on the effect of the environment on the community. For example, a steep plot signifies a community with high dominance, while a slow decay implies high evenness. These shapes constitute null models against which observational data can be challenged. Thus, their systematic evaluation and characterization can assist in understanding microbial communities which remains a major issue in ecological studies.

5.1.2. Overview of rank-abundance distribution (RAD) curves

RADs plot species in decreasing order of their abundance along the x-axis, while the abundance of each species is displayed on the y-axis (often $\log_{10}$ or $\log_2$ transformed, to emphasize the rare species part of the distribution). Mugurran (2004) listed several advantages of using RAD curves over other methods. For instance, they

(a) clearly display contrasting patterns of species richness,

(b) highlight differences in the evenness amongst communities under study,

(c) can effectively illustrate changes through succession or following an environmental disturbance.

To seek general applicability and a better understanding of community organization, ecologists have applied various quantitative models to species abundance datasets (Barange and Campos, 1991). The most common of which are described below:

(a) *Log series model.* Initially proposed by Fisher et al. (1943), log series pattern will occur when species arrive at an unsaturated habitat at random intervals of time (Boswell and Patil, 1971; May 1975). It is common in communities which have low richness and predicts an exponential SAD.

(b) *Log-normal model.* Proposed by Preston (1948, 1962), log-normal distributions are common in communities with a few very common together with many rare species (i.e., highly right-skewed). It is the most commonly observed distribution in large assemblages and is usually considered the undisturbed default for most communities (Ulrich et al. 2010). As a community is disturbed, it tends towards the geometric series.

(c) *Geometric series model.* This was proposed by Motomura (1932), and is predicted to occur when species will arrive at regular intervals of time and will each occupy a fraction of the available niche space. It assumes that the early arriving species can pre-empt resources and can become dominant in abundance and thereby limit late comers (resulting in a steep dominance-distribution plot). It tends to hold for communities with low richness and where there is only one or a few dominating environmental factors (e.g., light limitation for understory plants, or soil nutrient limitation for desert microbes).

(d) *Broken stick model.* This was proposed by MacArthur (1957) and is common in those communities where most species are almost equally abundant. Hence, resource is partitioned more equitably and no severe dominance is possible resulting in a less-steep dominance-diversity plot. They are typically found in narrowly defined communities of closely related species.

These different models have been developed to describe species abundance data and can be broadly divided into statistical models and mechanistic models. The different forms of models result from

different processes. The advantage of statistical methods lies in that they enable the parameter of the distribution to be used as an index for biodiversity, which further facilitates the comparison of two communities. However, when the goal is to explain rather than describe, the biological or theoretical model is useful. Biological models can be based either on the assumption that an ecological community has a property called niche space that is divided among the species that live there, or on the assumption of neutrality (that is, the absence of niches)

5.1.3. Commonness and rarity

In previous sections, it has been mentioned that an ecological community is characterized by a few abundant and most rare species. Hanski (1982) proposed a “general rule” of nature where the common species are widely distributed and rare species are restricted. However, the boundary between the common and rare is a relative concept and depends on the scale of investigation. Also, sampling methodology may have a large impact on the perception of rarity. Rabinowitz and her colleagues (Rabinowitz 1981; Rabinowitz et al., 1986) proposed that a species rarity status is a function of three characteristics- geographic distribution, habitat specificity and local population size. Later, Gaston (1994) proposed that rare species as those that fall in the lower quartile in an assemblage. Likewise, the upper quartile can be used to identify common species (**Figure 5.1**), although this approach de-emphasizes the proportion of low abundance species in an assemblage (Maina and Howe, 2000). In addition, this quartile criterion may mask the differences in the preponderance of rare species in different assemblages. However, it provides a starting point to subsequently divide the species as rare (tail, hereafter) and common (head, hereafter) (**Section 5.2.5: defining head and tail species**).

In recent years, molecular methods for microbial community analysis have provided a new understanding of species distribution and diversity. The advent of culture-independent methods have created new opportunities to understand the genetic diversity, population structure, and ecological roles of communities of microorganisms. Perhaps molecular data has provided us with the greatest wealth of information, but the biggest problem has been to delimit species based on these sequences (single locus molecular data). A lack of suitable universal criteria to delimit Operational Taxonomic Units (OTU) the cluster of sequences roughly corresponding a microbial 'species' has remained a big challenge. This is because species are well separated in cases, like those of most multicellular organisms, where population size is small and birth rate is low. When organisms that are genetically distinguishable are morphologically very similar, or vice-versa, the correspondence between OTUs and species is more problematic. On the other hand, next generation DNA sequencing techniques have transformed microbial ecology. In particular, they allowed us to answer important questions such as what are the drivers of the enormous genetic and metabolic diversity in an environment by providing

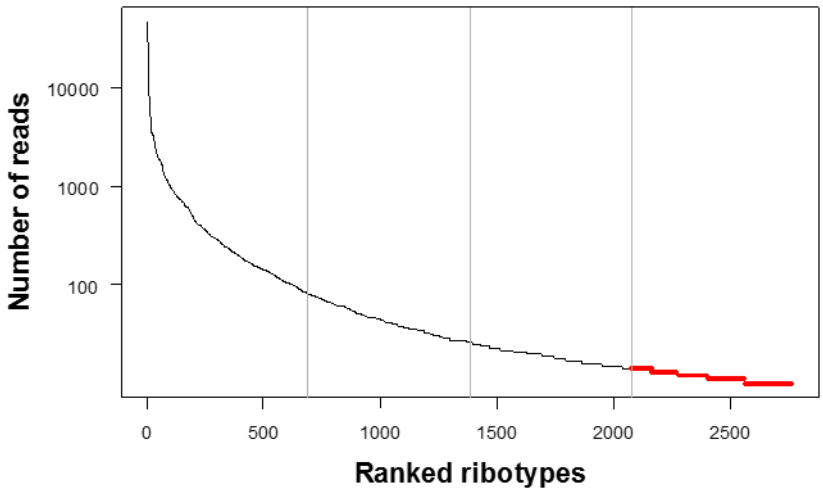


Figure 5.1 Delimiting rare species. Gaston (1994) defined rare species as those that fall in the lower quartile (in terms of proportion of species) of the species distribution model. Likewise the upper quartile can be used to identify common species.

a coverage of microbial diversity two-three orders of magnitude higher than morphology-based methods. In general, the dominant components of microbial communities always mask the detection of low abundance microorganisms which constitute a highly diverse “rare biosphere” in almost every environmental sample (Lauber et al. 2009). This “rare biosphere” is largely unexplored and offers a potentially inexhaustible genetic reservoir that is compatible with the ‘everything is everywhere, but the environment selects’ credo.

5.1.4. Power-law distribution

Despite their potential importance and wide occurrence in biological and ecological systems, power-law distributions remain little explored in ecological studies. In community ecology, many phenomena are characterized by a sudden production and slow loss or vice-versa, which is expected to lead towards the emergence of power-law structures. In general, they are expected in a system that are repeatedly forced away from equilibrium conditions (Barenblatt, 1996; Bak et al., 1988).

In general, power-law relationships are defined as the relationships where some quantity can be expressed as some power of another. It can be expressed as,

$$y = \beta x^{\alpha},$$

where, α is the power-law exponent. The distributions characterized by power law functions will appear as a straight line in log-log plots. It should be stressed that, when a distribution exhibits a power law-like shape, it can do so only over a finite range of event sizes, either bounded between a lower and an upper cut-off, or above a lower threshold, i.e., only in the tail of the observed distribution. RADs often display a fast exponential decay in the tail indicating the distribution with tails to be a power-law distribution.

Properties. There are two main characteristics associated with power-law relationships that set apart their theoretical and empirical importance. First, they display invariance under scale change, i.e., are scale-invariant (e.g., Sornette, 2000; Stanley et al., 2000; Gisiger, 2001), and second, their universality. Owing to their properties, the analysis of power-law relationships can help in identifying the existence of general/universal principles within ecological systems (Marquet et al., 2007).

Power-laws in ecology. The property of scale-invariance makes them very well suited for the study of ecological systems, which show variability at different temporal, spatial and organizational scales such that there is no single ‘correct scale’ for their analysis (Levin, 1992). In the past two decades, many studies have attempted to fit ecological data to diagnose ecosystem complexity (Miramontes 1995; Bak, 1996; Keitt and Marquet, 1996; Rhodes et al., 1997; Ferrier and Cazelles, 1999; Gisiger, 2001; Roy

et al., 2003; Pascual and Guichard, 2005). For instance, Li (2002) studied the macroecological patterns of pico- and nano-phytoplanktonic communities and reported that their total abundance was related to assemblage mean cell size according to the 3/4 power-law of allometric scaling in biology. However, the exponent has been reported to be closer to –1 when analyzing species in more than one trophic level (de Boer and Prins, 2002; Cohen et al., 2003). Passy and Legendre (2006) found that the behavior of higher taxa richness is a function of species richness that conforms to power-law. All these studies have interrogated biomass-size spectra, metabolic spectra, and population growth rates to resolve ecological complexity (Marquet et al. 2005). In recent years, power-laws in population fluctuations have also been the focus of research. Instead of focusing on differences among species in a comparative frame, these studies seek to separate general patterns that are invariant across taxonomic groups (Seuront, 2010).

The power-law behavior has been claimed for many distribution studies, for instance, species-area relationships (SARs) focusing on abundance distributions (Preston 1948; May, 1975), the allocation of individuals (Sizling and Storch, 2003; Plotkin et al., 2000), or population dynamics (Hubbell, 2001; Durrett and Levin, 1996). SARs are commonly described as a power-law function with a scaling exponent of ¼ (Lomolino, 2000; Schoener et al., 2001). Martín and Goldenfeld (2006) argued that power-law SARs are a robust consequence of a skewed species abundance distribution (SAD) resembling a log-normal with higher rarity. Based on mathematical relations, Irie and Tokita (2006) suggested a common mechanism for SADs and SARs. Using data on marine fish communities, Ferriere and Cazelles (1999) reported that intermittent rarity in communities of interacting species is governed by a well-defined -3/2 power-law.

A recent metagenomics analysis of marine phage communities suggested that the long-term decay of isolated phage populations follows a power law (Hoffmann et al., 2007; Edwards and Rohwer, 2005). These study showed that molecular data can aid in unveiling ecological complexity of phage communities. Other studies on microbial communities have shown exponential tails (Goldenfeld, 2013). It certainly would be interesting to investigate whether diversity in ecological systems can be explained by simple laws or principles. Power-law has emerged as a mathematical (and statistical) descriptor of the patterns in nature. Different communities, drawn at a global spatial scale, can be interrogated to see if there exist any universal patterns. If so then it would be equally interesting and important to see if it can be fitted with a power function.

5.1.5. Structure of the study

Microbial eukaryotes on Earth actively influence the functioning of the Earth system, however the vast majority are still uncultured and uncharacterized. The introduction of molecular-based, culture-independent techniques have facilitated new insights into the diversity of microbial eukaryotes and has assisted investigations of their global diversity. Using Illumina sequencing-based profiling (de Vargas et al., 2015) of the samples collected during the *Tara* Oceans expedition (Karsenti et al., 2011), a global community dataset was obtained. These data sets are unique in that they offer a view on marine ecosystems at different orders of magnitude in size. A total of 161 samples from sub-surface layer of the ocean, spanning four size classes, i.e. 0.8-5 µm, 5-20 µm, 20-180 µm and 180-2000 µm, were selected for this study. The underlying objective of this study was to study the organization of the sub-surface communities of protists to develop an understanding about the processes controlling it.

Different processes might be occurring at the same time even within a single level of observation, for instance, rare and common species might be subjected to different structuring processes if they are differently integrated in the ecosystem. To understand the varying impact of processes on the commonness or rarity pattern, the first step is to delineate rare and common ribotypes. In the present study, the variation in the total number of rare ribotypes across different sampling sites was studied.

In the next level, I attempted to develop a framework that can be used as a potential tool to identify and classify structures in marine ecosystems and also to infer the underlying processes that generate the observed patterns. Rank abundance curve was obtained for each sample and its shapes, especially of the tail, was studied. The characteristic shapes of the tail of RAD curve can be used to hypothesize its origin. For instance, in phytoplankton community distribution, it can be speculated that mixing or changing nutrients and/or zooplankton concentrations will alter its distribution and intensity to the extent that they will affect the characteristic exponents. Thus, the identification and the classification of the exponent of the tail could allow one to relate its change to some feature of the environment.

In the present study, I plotted RADs for all the protists in a globally distributed set of *Tara* Oceans surface samples and found a long-tailed distribution that appears to follow a power-law behavior. Further, a statistical approach was used to elucidate the nature and extent of microbial eukaryotic diversity. Li and co-workers (2012) have demonstrated that the variation of diversity within low abundant taxa cannot be sufficiently quantified with standard ecological diversity indices. This motivated me to see whether the slope of the tail fitted by power-law can be used as a diversity estimate. In brief, this work explores (i) the shape of the tail of RADs for marine protist communities, (ii) the range of the power law exponent of the fitted tail, (iii) commonness and rarity patterns. In

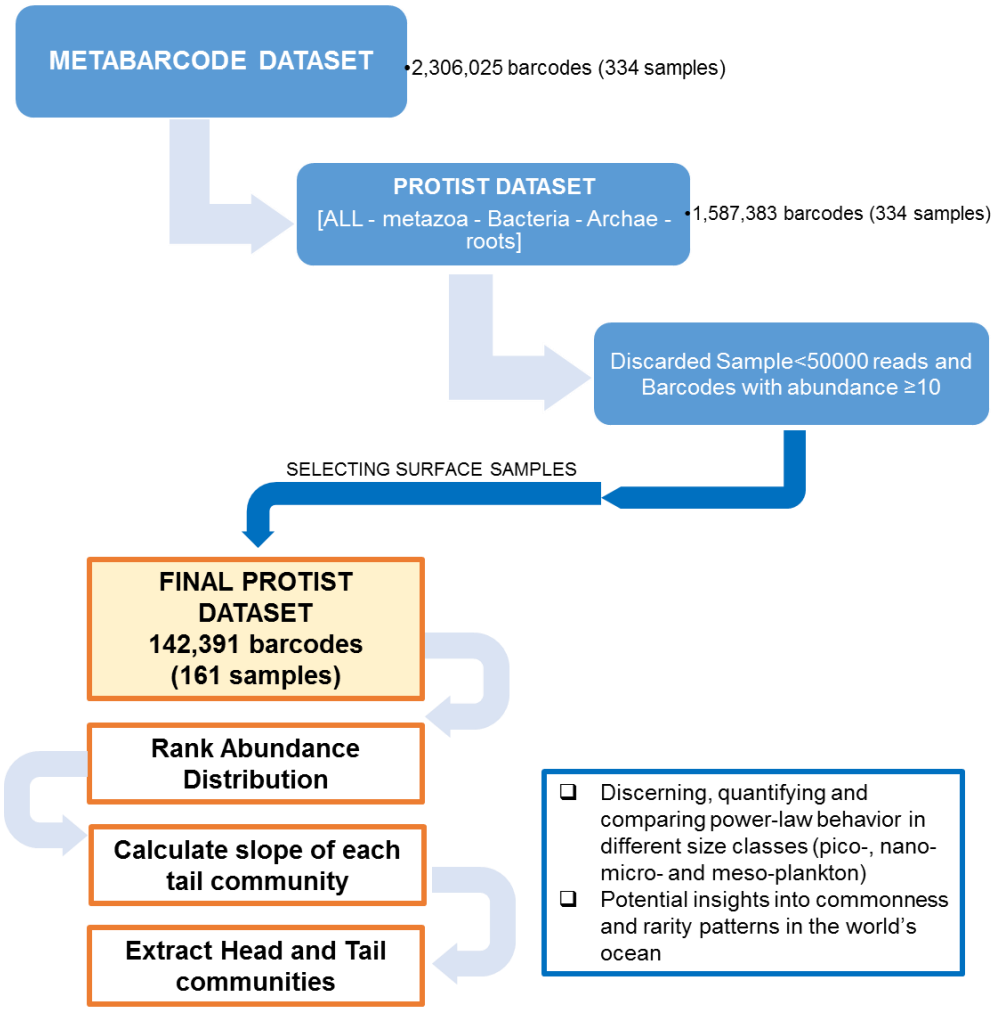


Figure 5.2. Protist dataset used in the study.

addition, we assessed the occurrence of a ribotype across sampling sites to understand the extent to which a rare ribotype remains rare in space (i.e., intermittent rarity).

5.2. Materials and methods

5.2.1. Protist dataset

Tara Oceans has provided an unprecedented opportunity to examine microbial diversity within and across the oceans through Illumina sequencing-based profiling of the hypervariable V9 region (or ribotypes) of 18s rDNA gene sequences from 46 stations across six oceanic provinces (For detail, see Chapter I). For the current study, we selected those ribotypes that were assigned to one of the protistan lineages from this comprehensive dataset. On this, we further applied a low-abundance filter, which discarded ribotypes whose abundance (total reads) did not exceed 10 reads in a sample, to avoid fluctuation due to sampling artifact or sequencing error. Thus, for the four size classes (pico-nano, nano-, micro- and meso-plankton), a total of 161 surface samples (132,187,241 reads; 137,249 unique ribotypes) were analyzed. The workflow is presented in **Figure 5.2**.

5.2.2. Delineating rare ribotypes in the world's ocean

There are two alternate ways to look at patterns of commonness and rarity, i.e., amongst samples and within a sample. Firstly, to study the pattern amongst samples, a rank was allocated to each ribotype in decreasing order of their abundance and mean relative rank was computed. This was done for each size class. The occupancy, i.e., the number of stations in which the ribotype was found in, was plotted against its mean relative rank.

Secondly, to delineate head (common) and tail (rare) ribotypes within a sample, a fixed threshold of 60% was used. More specifically, the first 40% of the ribotypes were tagged as head and later 60 % as tail, thus each community was divided into head (common) and tail (rare) sub-communities. The community dissimilarity was calculated using Jaccard distance and communities were then clustered based on average linkage to see if the tail is a result of sequencing error. Also the ribotype composition of tail communities were studied to explore how these vary across stations (locally) / oceans (regionally).

5.2.3. Fitting power-laws to the community data

Many ecological models can be written as: $y = f(x, P) + \varepsilon$, where y is a vector of *n* measurements of a response variable, x is a vector of predictor variables, P is a vector of p unknown parameters, and ε is a vector of errors, currently called residuals. To apply the maximum likelihood method, a probability

distribution (often normal distribution) is assumed for the residuals.

Maximum Likelihood Estimation (MLE). Goldstein et al. (2004) stressed that methods for determining the exponent of a power-law tail by graphical means, often used in practice, provide biased and inaccurate estimates. MLE offers a mathematically sound alternative to graphical methods to produce more accurate and robust estimates of a power-law tail exponent. MLE was first developed by the famous geneticist and statistician R. A. Fisher. It has been accepted as a standard statistical technique to estimate model parameters. As the name implies, MLE proceeds to maximize a likelihood function, which measures the agreement between the model and the data. For each data point, one has a function of the distribution's parameters. The joint likelihood of the full data set is the product of these functions. This product is generally very small indeed, so the likelihood function is normally replaced by a log-likelihood function. MLE has been recommended on practical grounds as the most popular estimation technique in statistics owing to its four theoretical properties, i.e. consistency, asymptotic normality, asymptotic efficiency and asymptotic invariance.

Computing the power-law exponent using maximum-likelihood. We attempted to establish a method for identifying the power-law regime and for estimating the exponent of the power-law regime. The R function, *stats4,* was used to fit the RAD with MLE. The steps used were as follows:

Step 1. For each community, ribotypes were ranked in decreasing order of their abundance.

Step 2. An initial set of "tail" ribotype sub-communities were chosen based on Gaston's quartile criterion (1994). The lowest-rank ribotype of the tail was stored (q).

Step 3. An initial fit to a power-law (linear function in log-log scale) was performed, using MLE, with all ribotypes of the sample under study.

Step 4. The same procedure was repeated on progressively smaller sub-communities, obtained augmenting of one the minimal rank represented. In this way, the fit is performed on parts of the RAD that progressively exclude the more abundant ribotypes.

Step 5. The log-likelihood function for the MLE of the linear fit was stored.

Step 6. Repeat step 4, until the ribotype ranked "q" was met.

Step 7. Out of the set of candidate models for each community, the model with the maximum log-likelihood value was selected and the set of parameters were reported. The standard error of the linear regression was also reported as a measure of precision to the slope estimate.

5.3. Results

The results from this study are divided into two major sections: (i) potential insights into commonness and rarity patterns of protists in the world's ocean, and (ii) discerning, quantifying and comparing power-law behavior in different size classes (pico-, nano-, micro- and meso-plankton).

5.3.1. Potential insights into commonness and rarity patterns of protists in the world's ocean

For each size class, the occupancy of all ribotypes under study were plotted against mean ranks (**Figure 5.3**). Each quadrant was characterized as follows:

Quadrant I: high mean relative rank and high occupancy,

Quadrant II: high mean relative rank and low occupancy,

Quadrant III: low mean relative rank and low occupancy,

Quadrant IV: low mean relative rank and high occupancy.

Of these, we found that the majority (~82-90%) of the ribotypes were seen in the third quadrant (i.e., low mean relative rank and low occupancy), indicating that the majority of the ribotypes in the head are not cosmopolitan, in general. On the other hand there were only a few ribotypes in the first quadrant (i.e., high mean relative rank and high occupancy), indicating that rare ribotypes are not cosmopolitan. The overall shape of the distribution was very different for the smallest size fraction in comparison to the others (**Figure 5.3**). Ribotypes with high mean relative rank and high occupancy were seen only in the smallest size fraction. The taxonomic composition of each quadrant revealed that the overall composition of quadrant III (QIII; with majority of ribotypes) varies greatly among the four size classes (**Figure 5.4**). For the smallest size class, QIII was highly diverse, whereas it was dominated by Collodaria in the largest size class.

A "head" and "tail" community for each station was obtained using a fixed threshold (40% and 60% in log-log scale, respectively). This threshold was used to obtain the head and tail sub-communities for each sample under study. Most of the ribotypes were seen in the "head" community in one station and in the "tail" community in other stations. This substantial overlap, i.e., the ribotypes appearing in both head and tail, suggested that there exist only a few ribotypes that were exclusively heads or tails, favoring the view of "intermittent rarity" (Ferriere and Cazelles, 1999). The community dissimilarity for "head" and "tail" were calculated using Jaccard distance and the dendrograms (**Figure 5.5**) illustrated that the "tail" (rare) communities clustered separately from the "head" (common) of the same RADs. This would not be the case if the tails were the effect of sequencing errors, in which case the probability that the same error occurred repeatedly in two different samples (even with a similar composition) would be minimal in the absence of massive sequencing biases. Thus, the study of dissimilarity

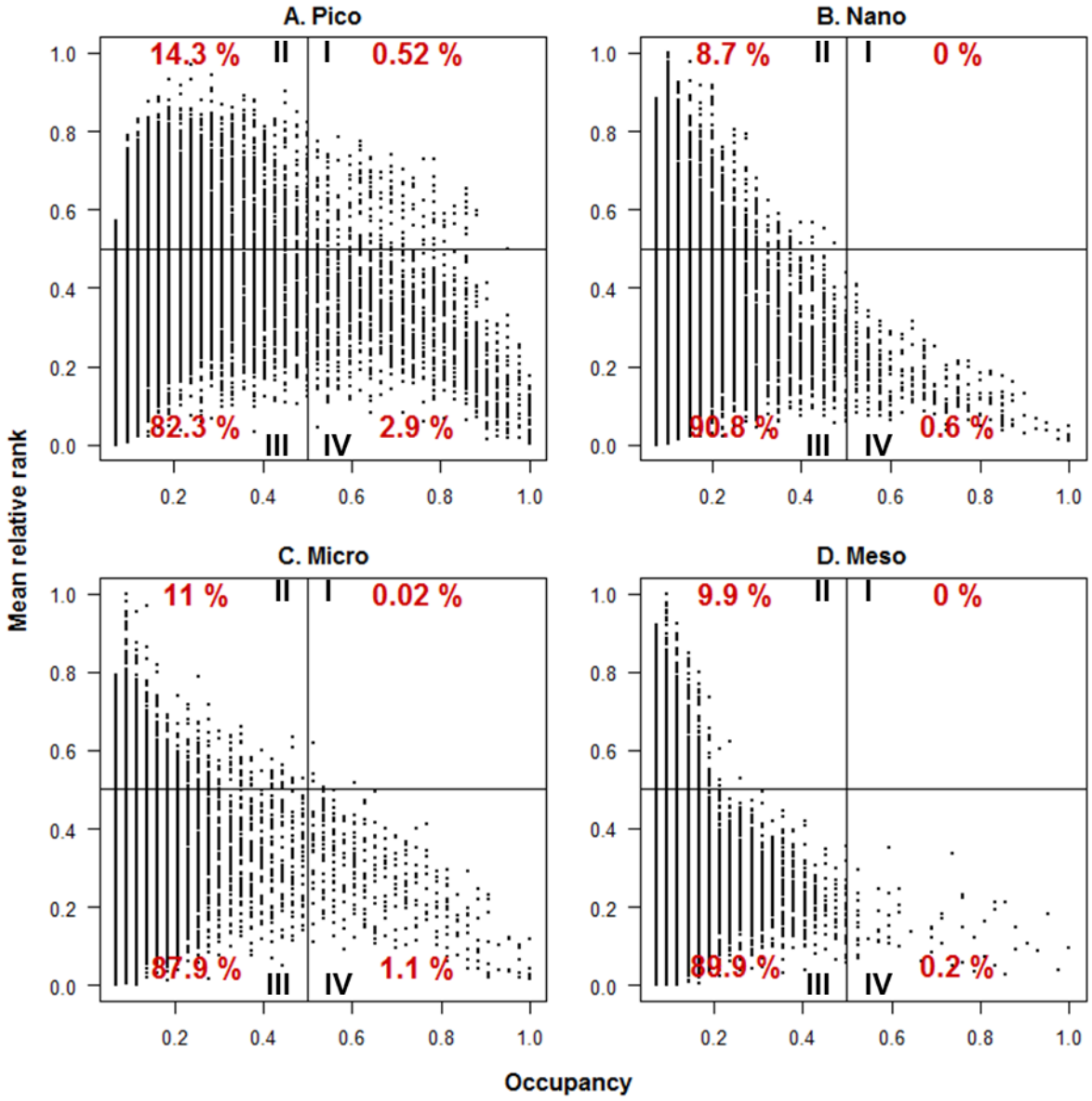


Figure 5.3. Occupancy vs mean relative rank. (A) Pico (0.8-5 µm), (B) nano (5-20 µm), (C) micro (20-180 µm) and (D) meso (180-2000 µm). Occupancy correspond to the number of stations in which a ribotype is seen, relativized by total number of stations for each group under study.

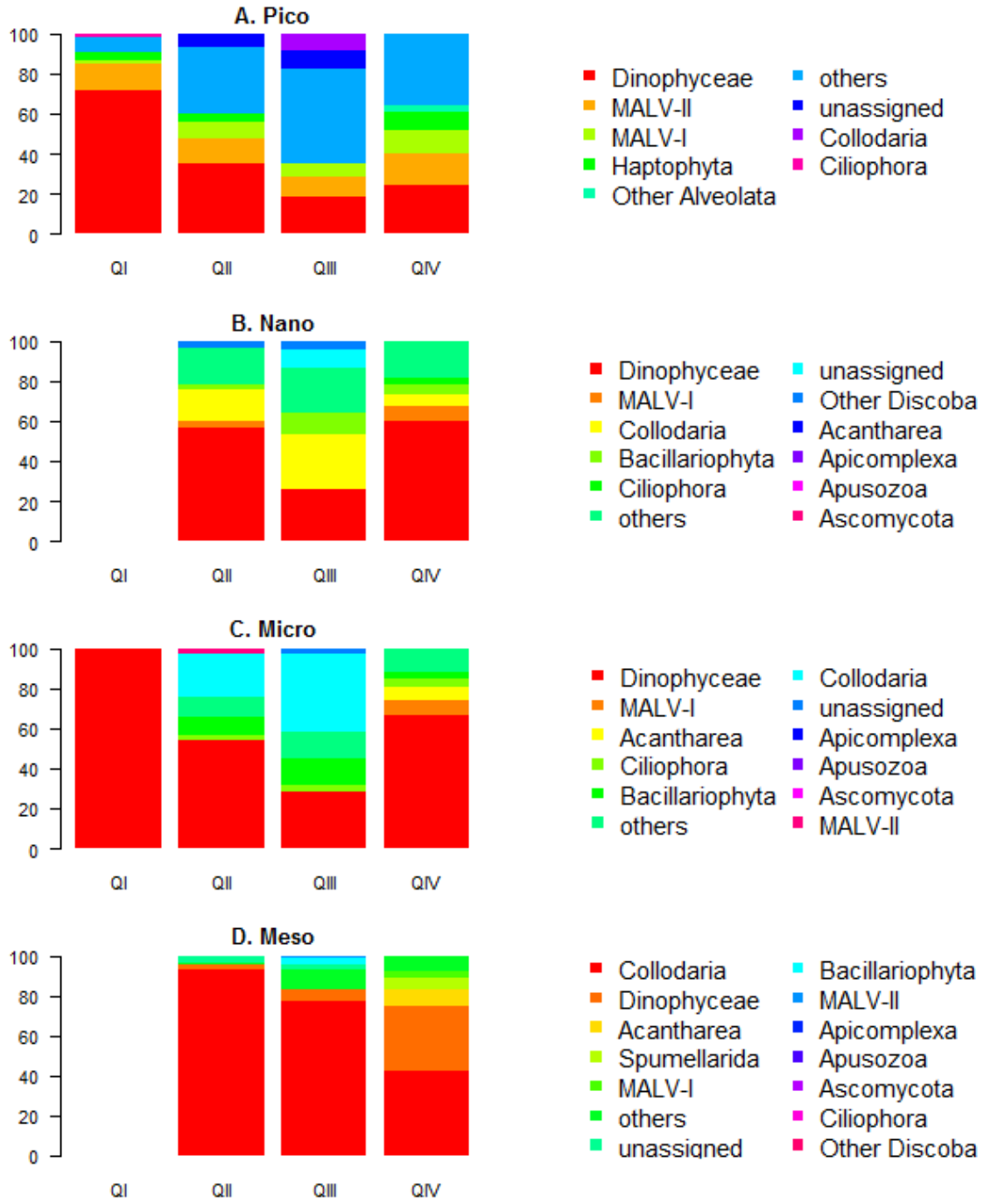


Figure 5.4. Taxonomic composition for each quadrant. (A) pico (0.8-5 µm), (B) nano (5-20 µm), (C) micro (20-180 µm) and (D) meso (180-2000 µm).

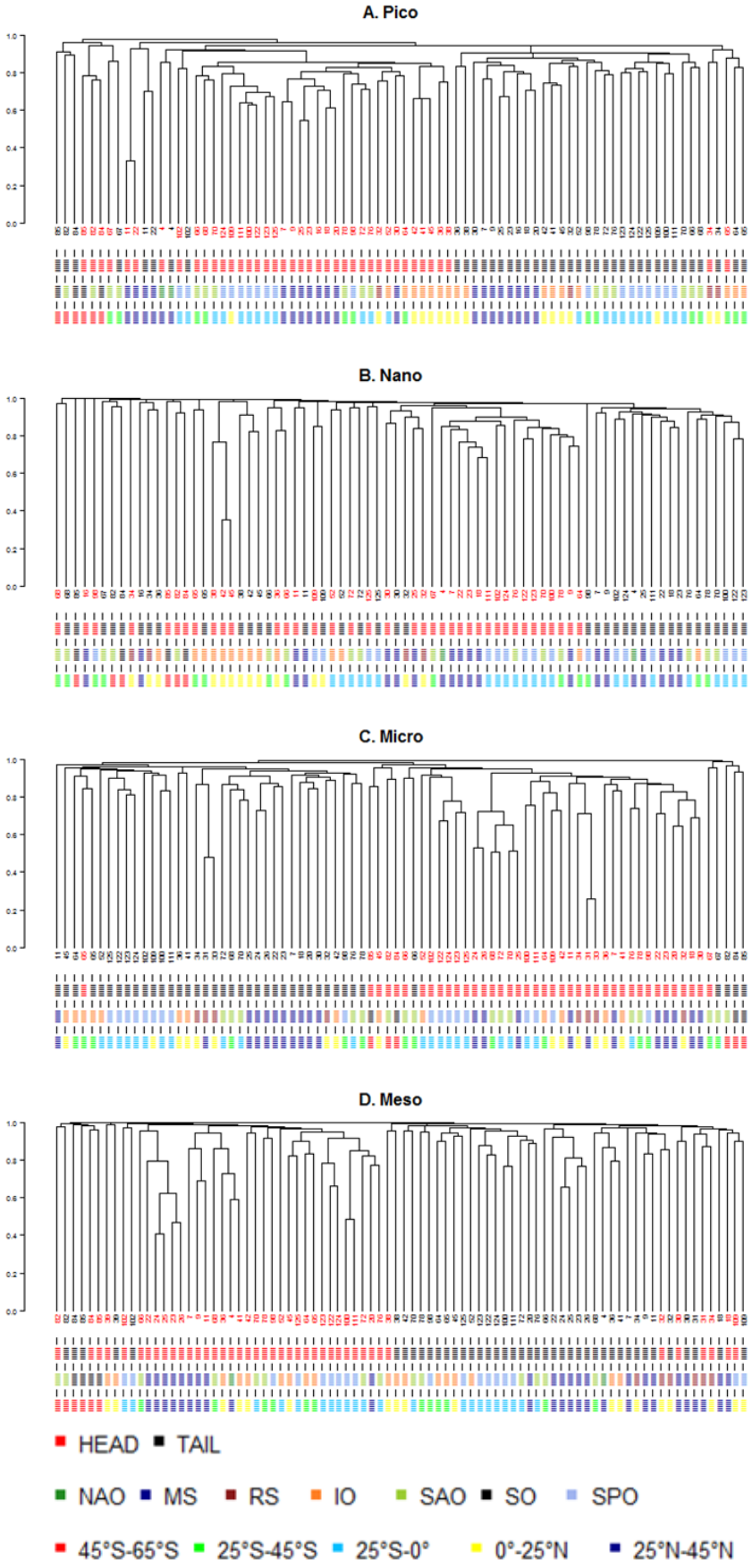


Figure 5.5. Hierarchical clustering of head and tail communities. (A) pico (0.8-5 µm), (B) nano (5-20 µm), (C) micro (20-180 µm) and (D) meso (180-2000 µm).

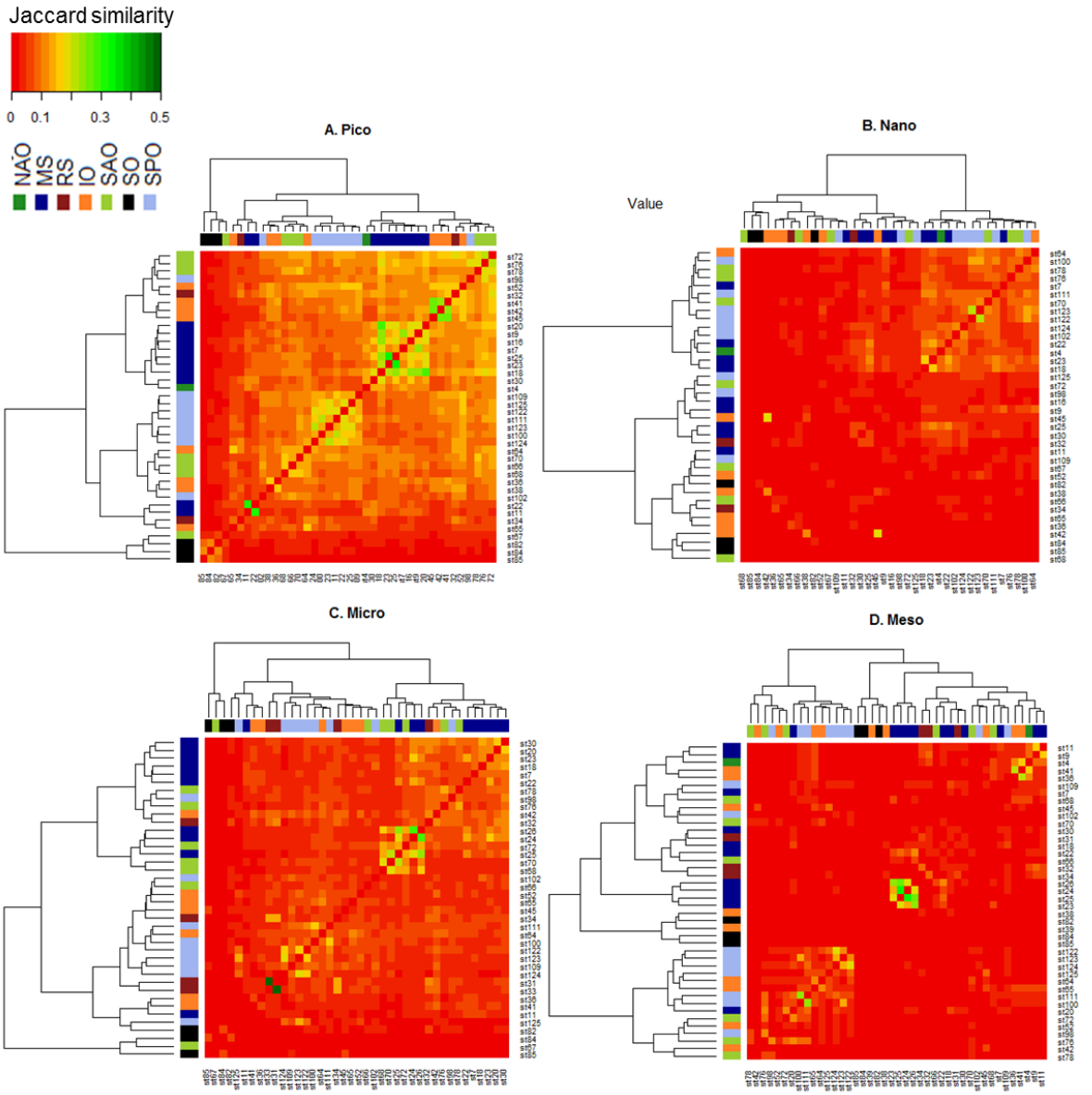


Figure 5.6. How similar is the tail across stations? (A) Pico (0.8-5 µm), (B) nano (5-20 µm), (C) micro (20-180 µm) and (D) meso (180-2000 µm).

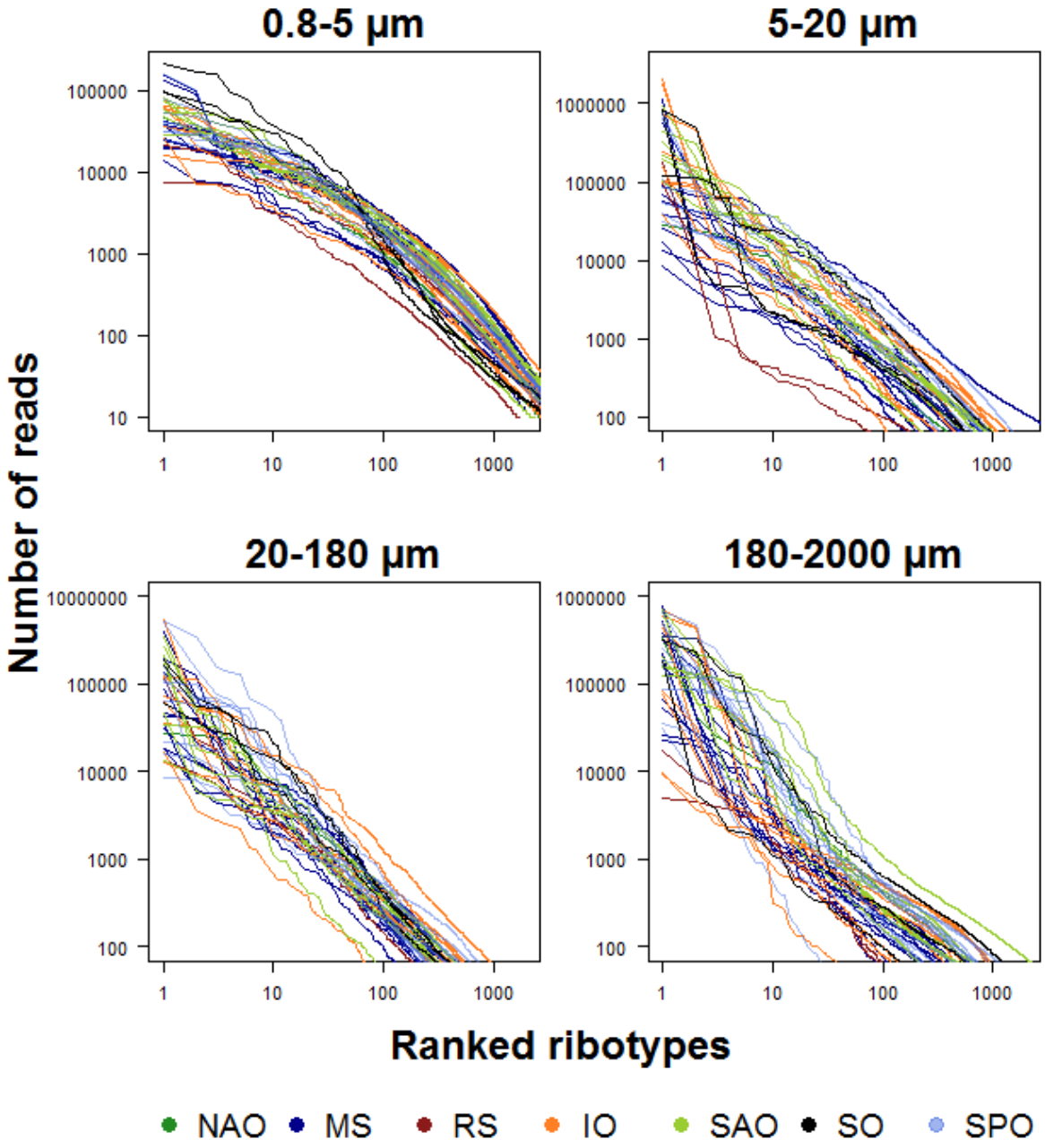


Figure 5.7. RAD grouped by size classes. Each curve represent a station color-code by the oceanic province.

between common and rare communities suggest that the tail does not appear to be due to sequencing errors. Moreover, the same conclusions are maintained at coarser taxonomic resolutions, when ribotypes are clustered into swarms. The tail-communities were found to differ significantly across various sampling stations, as evident from the very low (<0.15) Jaccard similarity indices (**Figure 5.6)**, consistently with the observation of the paucity of rare cosmopolitan ribotypes.

5.3.2. Discerning, quantifying and comparing power-law behavior

5.3.2.1. Rank-abundance distributions (RADs)

For each community (one size class of one station), the RAD based on ribotypes was plotted on a log-log scale (**Figure 5.7**). Interestingly, on the first hand our data indicated that the tail appears to follow a power-law behavior for almost every protistan community of different sizes.

5.3.2.2. Power-law fit to the community dataset

In this study, I chose an initial subset of a rank ordered list (descending) of ribotypes (as tail; red curve in **Figures H5.1,** annex) in the lower quantile following Gaston's quantile approach. This limit was extended (blue curve in **Figures H5.1,** annex) to include those ribotypes which best fit a power-law model. The exponent of such a power-law was estimated using the maximum likelihood (MLE) approach (**Figures H5.1,** annex). The slope varies primarily between size, independently of the ecosystem studied and the sampling period (**Figure 5.8**). For the pico-protist fraction, the slope statistic was found with a mean value of -1.54 and median of -1.58 (range: -1.96 to -0.98). For the nano-protist fraction, a slope statistic was found with a mean value of -1.33 and median of -1.32 (range: -2.29 to -0.87). For the micro-protist fraction, the slope statistic was found with a mean value of -1.25 and median of -1.23 (range: -1.78 to -0.9). For the meso-protist fraction, the slope statistic was found with a mean value of -1.41 and median of -0.89 (range: -2.36 to -0.89). Interestingly, dropping a few outliers showed that the slope has very little variation and was in the range of -1.6 to -1.2 for all size fractions (shaded region, **Figure 5.8**). The power-law fit was more accurate in the pico- and micro size-classes, whereas it was not found to be a very convincing choice for the other two size-classes, probably due to under-sampling. Also, most of the outliers (mainly the large size-class) were the cases where the power-law was not a good fit, and all the outliers in the larger size fraction were from the low abundant communities, where a power-law may be potentially masked by under-sampling. Interestingly, all the communities in the Southern Ocean were outliers in the pico-communities for the power-law exponent, due to its large population size. In general, power-law was not a good fit for either the very abundant and low abundant communities. On one hand, the most stations with higher total abundance of ribotypes deviated toward the less steeper (less negative slope) and on the other hand the low

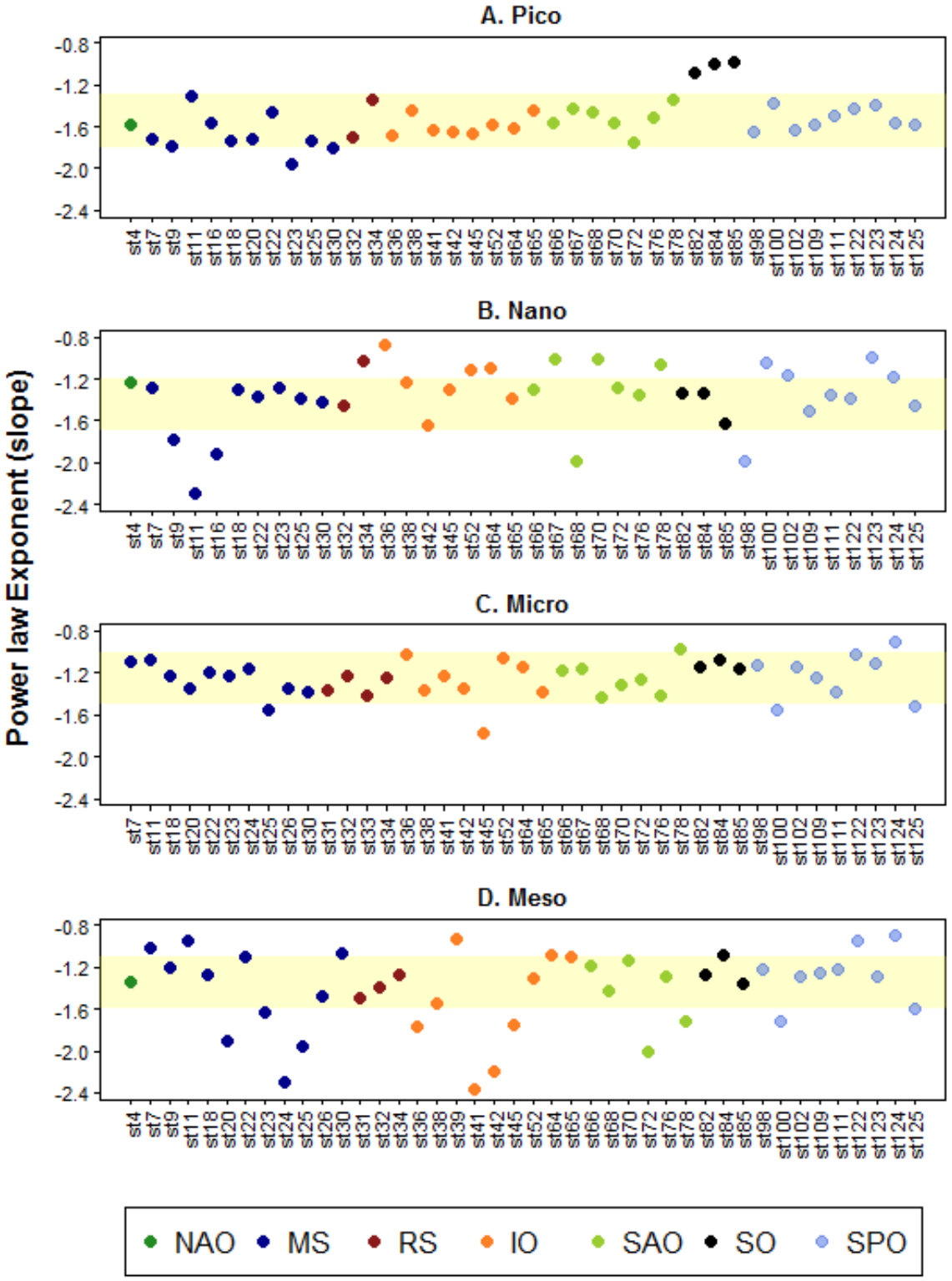


Figure 5.8. Variation in slope in communities from different sampling stations for each size class.

abundant stations exhibited a steep slope. The length over which the RADs were fitted with power-law was over 3 decades in most cases. For the pico-communities, the size of the fit centered around 3.5-3.75 decade class (**Figures 5.9**). Additionally, it appeared that the gradient of the fit was more similar within a station.

5.3.2.3. Variation in power-law exponent

Across size classes. We then explored whether general patterns in slope variation emerged across size classes (**Figures 5.10**). The meso-protist fraction showed the highest median slope whereas the pico-protist fraction exhibited the lowest median slope. Mann-Whitney-Wilcoxon test was used to test if the median slopes were significantly different for each comparison (**Table H5.1,** annex) and it was found that the median slope for the pico-community was significantly different from the other size classes. A remarkable difference was found in the pico-protist community with respect to other protistan size classes.

Across different oceanic provinces. For each oceanic province, the median slope of the tail showed remarkable variation for pico- and nano- communities. However, for bigger size classes comparatively less variation was seen. The median slope grouped based on the oceanic provinces showed that for the pico-community, the Mediterranean province has a very different slope, but for the rest of the size classes it was closer to the slopes of the other provinces (**Figure 5.11 and H5.2**).

Across different latitudinal bands. The median slope grouped according to latitudinal bands showed unique patterns (**Figure 5.12 and H5.3**). For pico- and nano- communities, the southernmost latitudinal band exhibited a different median slope in comparison to the others. Besides this exception the median slope was highly similar in almost all latitudinal bands.

5.3.2.4. Is there any correlation between slopes and other indicators of diversity

For each size class, the Shannon Diversity Index (SDI) and richness were compared against the estimated slopes. We found a significantly strong negative correlation between SDI and slope for the smallest size-class only (0.8-5 µm) (**Figure 5.13A-C**). For the richness, a significant positive correlation was seen for the micro community (20-180 µm) (**Figure 5.13B-C**). The pico-community showed an opposite relation to SDI (and richness) in comparison to the larger size-size-classes.

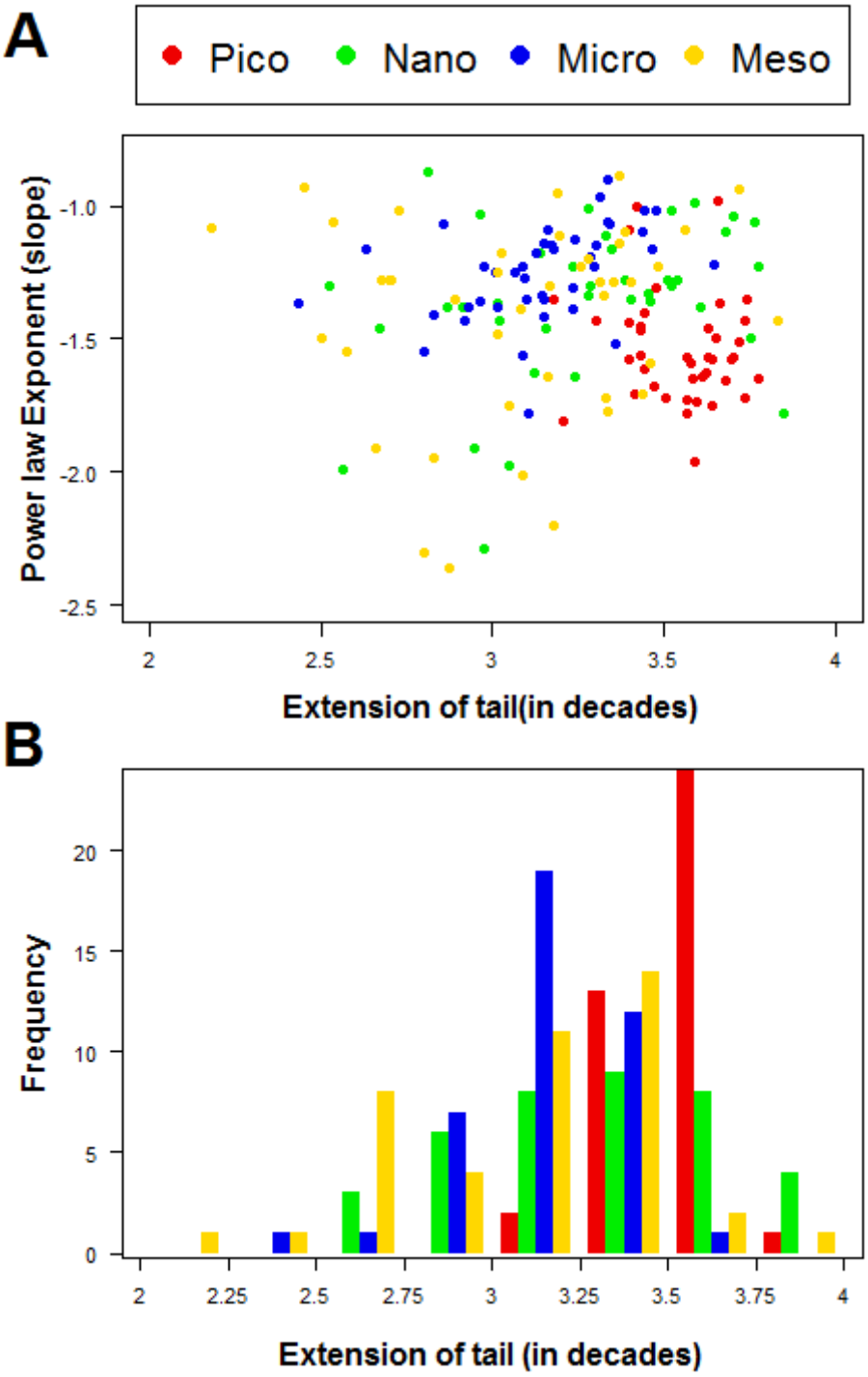


Figure 5.9. Size of the fit. (A) Scatterplot of the value of the slope vs the extension of the tail. **(B)** Frequency of computed slope exponents against each extension of tail corresponding to the size of fit.

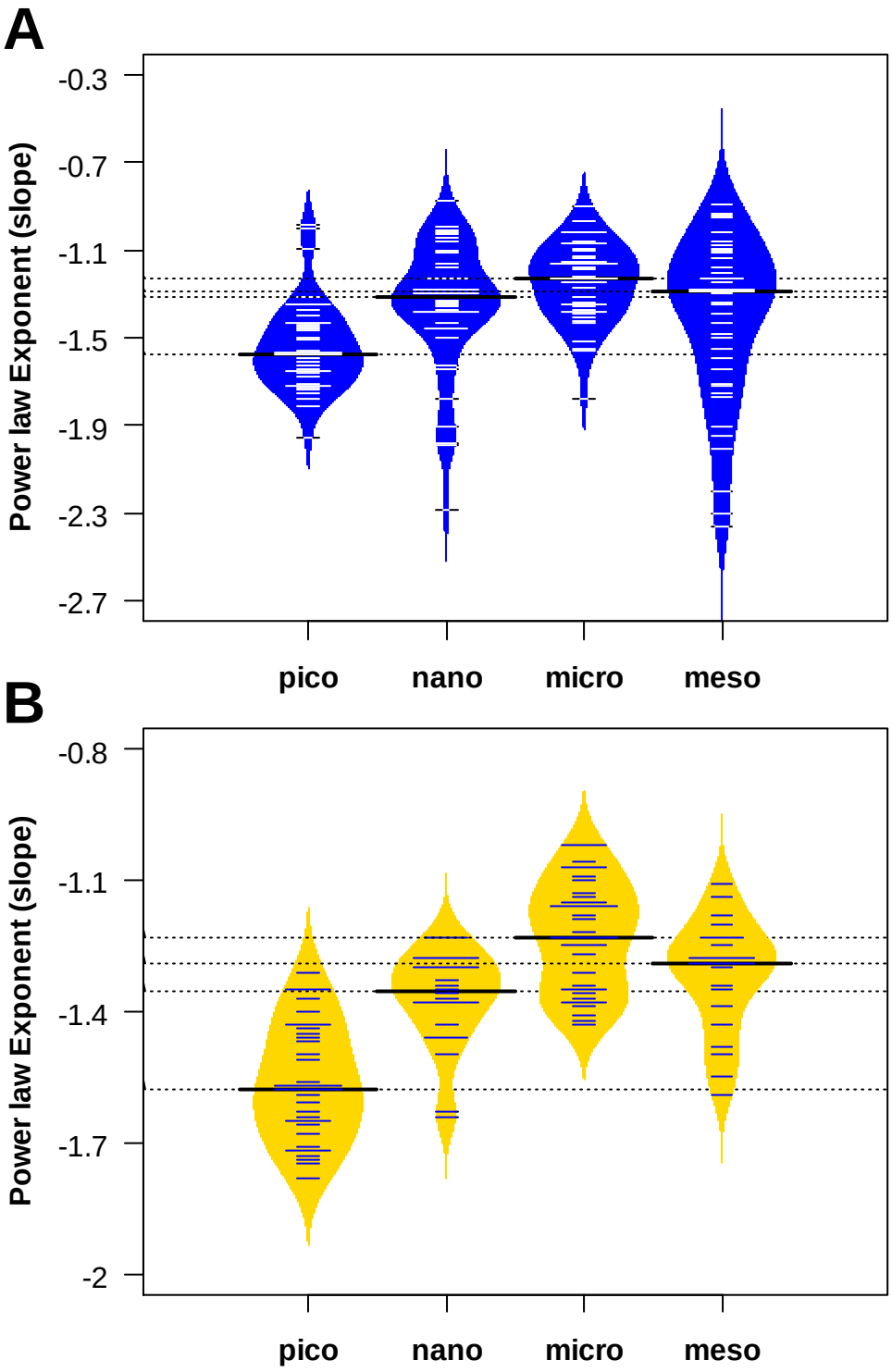


Figure 5.10. Boxplot of slopes clustered based on size. (A) For all communities, (B) without outliers (pale shaded area in Figure 5.9).The overall dashed line represents the median for each class.

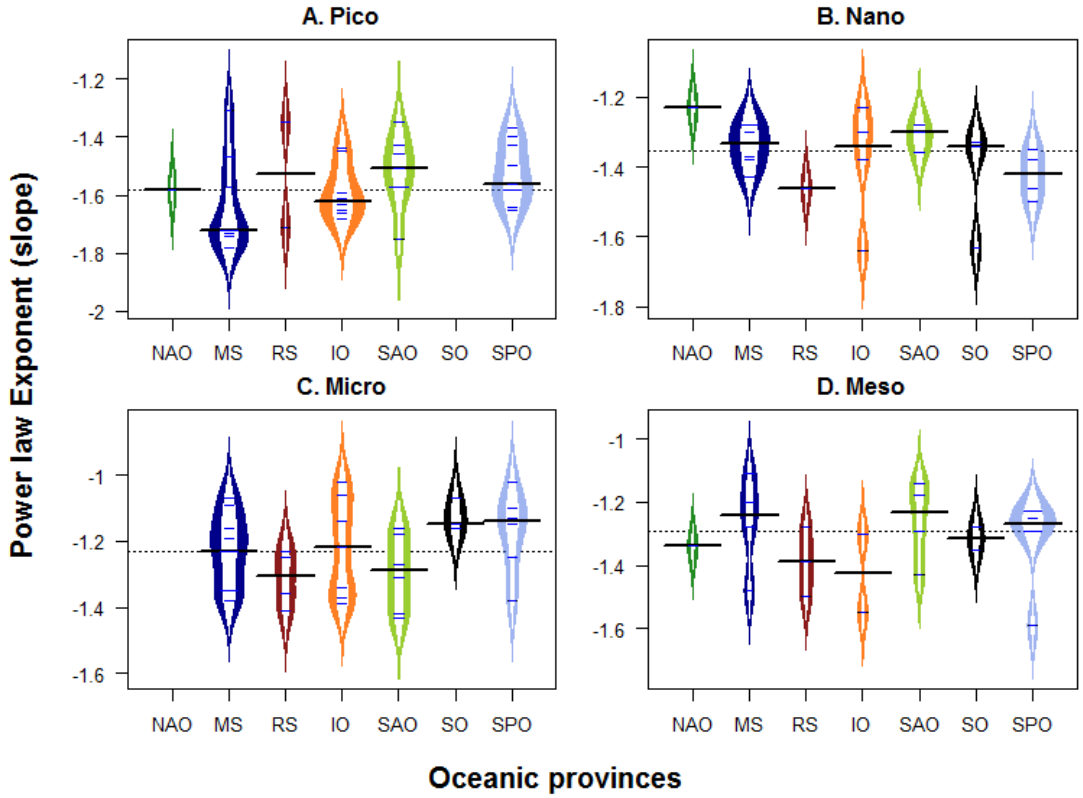


Figure 5.11. Boxplot of slope clustered based on oceanic provinces. Slope of the communities in the shaded portion in Figure 5.9 are used.

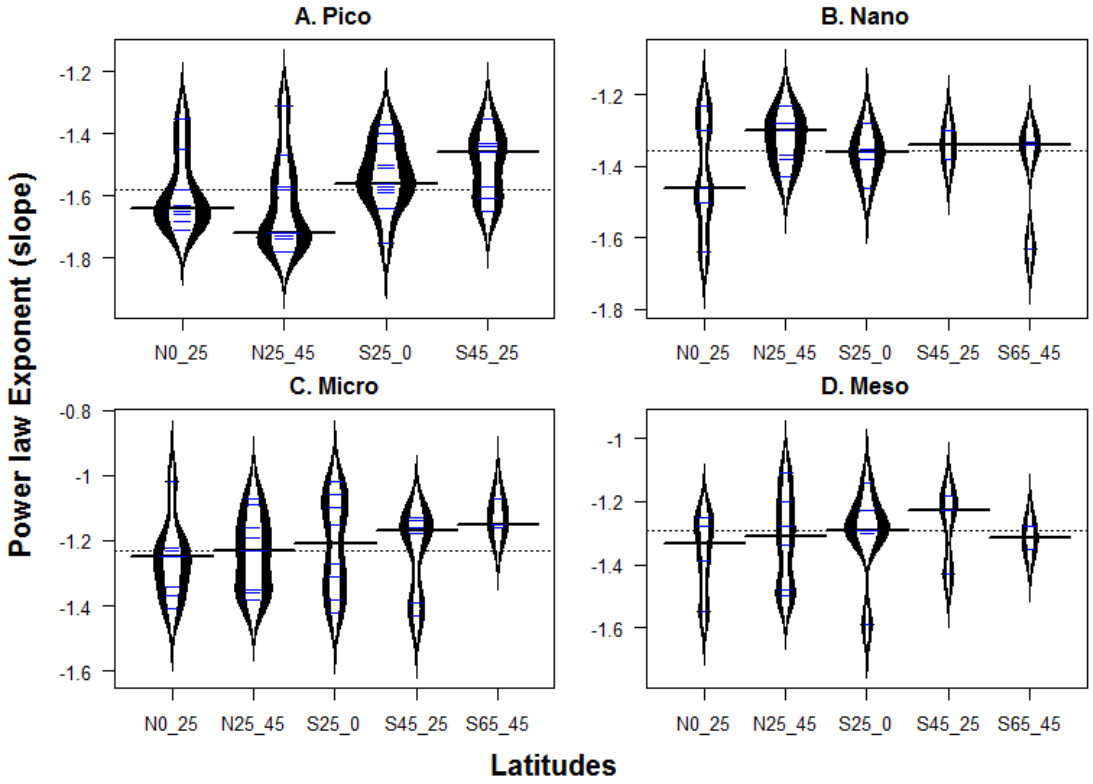
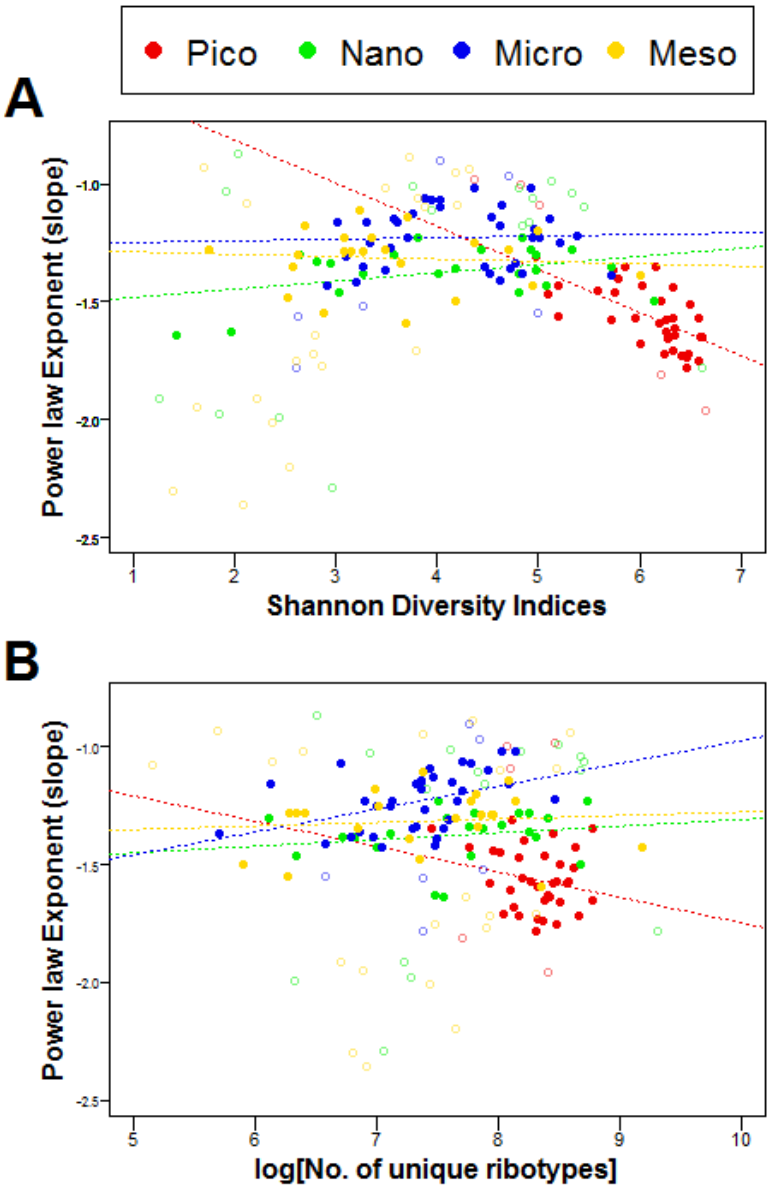


Figure 5.12. For all size fractions, boxplot of slope clustered based on latitudinal bands. Slopes of the communities in the shaded portion in Figure 5.9 are used.



C

Size class	Shannon vs slope		Richness vs slope	
	r	p-value	r	p-value
Pico	**-0.66**	**0.00**	-0.17	0.343
Nano	0.39	0.070	0.22	0.324
Micro	0.04	0.820	**0.45**	**0.007**
Meso	-0.08	0.734	0.08	0.599

Figure 5.13. Is there is any correlation between slopes and other indicators of diversity? Comparison of (A) Shannon Diversity Index vs slope, and (B) log(richness) vs slope. For each size class, the Shannon Diversity Index was compared against the estimated slope. Dotted line represents a regression line across each point. There is a closer relation between Shannon and slope than richness for all size classes. (C) Pearson's correlation coefficient. Open circle corresponds to outliers and were dropped from regression and correlation tests.

5.4. Discussion

Quantifying diversity is of central importance for the study of structure, function and evolution of microbial communities. The estimation of microbial diversity has received renewed attention with the advent of large-sale metabarcoding studies, considering that diversity observed in a sample tells us about the diversity of the community being sampled. Many independent studies, from terrestrial to marine, have reported that an ecological community is characterized by a few abundant and most rare species.

5.4.1. Potential insights into commonness and rarity patterns in the world's ocean

Rarity is common among ecological communities and is defined by two attributes, abundance and spatial distribution. An open question is why a particular species is common or rare. Another related question is: what are the potential ecological mechanisms that influence rarity or commonness? Kunin (1997) demonstrated that all species are rare everywhere. However, other studies have reported that some species are rare everywhere (Orians1997). A number of causes of rarity have been reported in literature (Gaston 1994; Kunin, 1997). The three major mechanisms include: ecological specialization, lack of dispersal, and historical contingency. To understand how rarity varies, one needs to define a "boundary" to distinguish rare from common. We chose to regard rarity as simply being the state of having a low abundance and/or a low occupancy (Figure 5.3). It should be emphasized here that rarity is influenced by the spatial scale at which the study is performed and that the categorization is made with respect to that particular spatial scale. Sampling artifacts is yet another issue which can inflate the number of ribotypes detected as rare. However, sometimes a rare species is not detected at all which leads to an underestimate of rare species.

The temporal and spatial dynamics of many populations involves intermittent rarity, that is, the alteration of extremely low abundance and short outbreaks. One of the major causes of this could be the result of competitive interactions within and between species. These intermittently rare species serve as weak invaders in fluctuating communities (Ferriere and Cazelles, 1999). On a temporal scale, Ferriere and Cazelles (1999) proposed that the intermittent rarity is governed by a well-defined power-law and that the scaling parameter (-3/2) is a universal feature of it. In brief, the dynamics of rarity have no characteristic time scale. At a global scale, the rarity and endemism (occurring only within a restricted area) are associated but not interchangeable concepts. This is well illustrated by the statement made by Kruckeberg and Rabinowitz (1985) that "the narrow or local endemic is the one that best fits the colloquial notion of rarity. However, the term endemism, in its classical geographic usage does not imply rarity or even small range". Ridley (1993) demonstrated that rarity may influence

evolutionary dynamics through its two aspects (i.e., small local population and small range size). Other studies have reported that rarity may have differing evolutionary roles at different spatial, temporal or comparative scales.

5.4.2. How plankton gets dispersed in random environments

In this study, I explored different properties of the common versus the rare occurrences, separating 'heads' and 'tails' of the distribution within every sample. This way, one can think of partitioning a community into common species that are adapted to the local environment (in the 'head', which will probably take an exponential form), and in rare species that may subsist in the background due to reduced selective pressure and immigration. Indeed, the expectations on its structure are different depending on the relation between the ribotypes of the 'heads' and 'tails'. Further, if the tail was generated by sequencing mistakes or by small-scale ('within species') radiation, one would expect the tail to always cluster together with the head; different samples would give rise to different tails. Such a result might also be the consequence of the tight relationship of rare species within communities dominated by abundant ones. If the tail reflected instead a 'rare biosphere' that is present in the background whatever the dominant species are, then the tails might cluster (**Figure 5.5**) together at a scale larger with respect to the heads. Heads would show a stronger biogeographical connotation, and have a stronger correlation with environmental variables; tails would reflect the relationship with the environment on a longer time scale, and thus be more uniform across spatial and temporal scales due to stirring. Still, one would expect that biogeographical domains would be visible, e.g., that the rare biosphere of the North Pacific would be separated from the rare biosphere of the Indian Ocean (**Figure 5.6**).

5.4.3. Power laws in ecology

Rank abundance distribution has been the method of choice to study species abundance distributions. The application of the power-law model, $\log(p) = \log(c) + z \log(a)$, to describing species distribution was first proposed by Gaston (1994) and Kunin (1998). It was observed that ribotype RADs do not appear to fit to one single distribution type which indicates that it should be explained as a superposition of RAD types rather than a single one. The tail of rare species was fitted independently of the rest of the distribution with one law, while the head of most abundant ribotypes may be explained by another law (and possibly a different one in blooming stations with respect to oligotrophic ones). The crossing-over to a different 'regime' for rare species may however be absent, or take different forms depending on the way the ecosystem is looked at. At the first instance, it was observed that ribotype RADs have a power-law tail (Zipf's law) in almost all samples. However, the distribution

of common ribotypes exhibited a variation in shape. The striking feature of the distribution of rare species is that not only the power law extends over several orders of magnitude (usually, ecological data do not allow such a resolution), but also that the exponent of all samples lies in the interval (-2 to -1; Figure 5.9), suggesting a common origin. Interestingly, it was found that the median exponent appeared to be different in different size classes, indicating a fatter tail with the increase in size. Such a common origin could be artifactual, either due to similarities in the error patterns of the amplification-sequencing, or due to sampling issues. But it could equally be due to biological and ecological features of the ecosystem, for instance the persistence of rare species in a 'neutral' life style. The power-law would be in this case explained by the ongoing evolutionary processes at the molecular level, that continuously create new variants by mutation and recombination, and by the fact that those variants are not strongly selected. Other possible sources of polynomial tails might be intermittent physical forcing by ocean turbulence or another complex ecological dynamics of the planktonic populations. It is not entirely clear how to tease apart these different potential explanations.

In the ocean, theoretical models classically predict that phytoplankton community structure is shaped heavily by ecological processes such as competition and dispersal. Both these processes are coupled to the peculiar physical properties of the open ocean such as advection and turbulence (Li, 2002). Competition is also likely to be a dominant process at short time scales, and in instances where the community experiences a bloom. Variants/species that are most adapted to a given environmental condition (typically defined by the water mass that contains the community) may quickly take over the community, thus resulting in a RAD very concentrated on a few classes (the blooming species and those that are directly connected to them because of mutualistic/parasitic/trophic interactions). The effect of competition may however be absent on the tail, representing rare classes that are completely out of the game. Such rare classes may persist in forms of dormancy, or take advantage of not having a direct competition with others. They may be carried around by currents and would be little affected by the ecological dynamics at a specific point in time and space. They would be broadly distributed, and they would correlate more weakly with specific environmental conditions.

5.4.4. Explanations of power-laws

The commonly used ecological indices for quantifying inter-sample diversity, Shannon and Simpson's, can perform well when approximating the microbial diversity of common taxa. However, each may fall short as a single complete measure when examining the numerous low abundant organisms that dominate the composition of many microbial communities. Since these indices are unable to capture enough of the low abundant taxa, I attempted to formulate a rank based diversity measure, i.e., the slope of the tail. But this statistic was found to be stable and, hence, anything that is almost constant

cannot be used as a measure of variation. On the other hand, its constancy is what makes it interesting if one looks for 'universal' laws.

With the aforementioned frequent occurrence of power-law distributions in biology and elsewhere, it seems natural to ask what are the explanation(s) for this widespread statistical feature in nature? Several reasonable mechanisms have been proposed that can possibly lead to a power-law behavior. Three of the utmost reported notions are (a) Out-of-equilibrium phase transitions and self-organized criticality (SOC; Bak et al., 1987), a property of dynamical systems that was proposed to explain the occurrence of complex phenomena. The term “self-organized criticality” emphasizes two aspects of system behavior. First, self-organization, to describe the ability of dynamic systems in the absence of control to develop specific structures and patterns. Second, criticality, to emphasize that a system stays at the border of stability and chaos; (b) highly optimized tolerance (HOT; Carlson and Doyle, 1999), a mechanism for complexity based on robustness tradeoffs in systems subject to uncertain (or evolving) environments. In evolution, it describes power-laws as optimal adaptations; and (c) theory of intermittent chaos (Pomeau and Manneville, 1979; 1980), projecting the intermittent chaos as the weak turbulent state in which the steady motion or the periodic motion is abruptly disturbed by random bursts, and that is found in many dynamical systems. For marine microbial ecosystems, which are far from equilibrium, power-laws may be explained due to turbulence and evolutionary branching. By getting an understanding of the principal mechanism, it is possible to identify the most relevant one for a given problem. This study is an attempt in this direction to elucidate the determinants of variability in species abundance, which has often been deemed as a central issue by ecologists. In this direction, the most important question one may ask is “whether those species that we presently regard as rare have also been rare in the past and are likely to be so in the future”.

To explain the power-law by ongoing evolutionary processes at the molecular level, it would be interesting to check if the power-law tail is still present when clustering is performed (either with swarms or OTUs, provided they are in sufficient number for 'seeing' the tail). In principle, if we assume that sequencing errors have a similar effect as point mutations, all sequences that are artificially generated should fall into the same swarm. The use of swarms might thus cut the tail of the distribution drastically by filtering out artefacts; although it would also filter out real standing genetic diversity. If the tails are maintained with the swarm classification, this could indicate that they are not artefactual. However, one should check whether the power-law is still the same when swarms are used instead of ribotypes. Another aspect is that power-law distributions are predicted by models of branching, so that they may be expected at the level of barcodes if the exclusion of newly arising variants is slow

enough with respect to the process of radiation. On the other hand, they may be expected at the level of 'species' (using OTUs/swarms as a proxy), if 'speciation' was a fast enough process.

To summarize, owing to their fundamental importance, the study of plankton biodiversity is a central element in ecological research. Many theories have been established to describe the distribution patterns and biodiversity at a global scale. This area of research has become important particularly due to the evolution of increasingly rapid circles under the influence of different actions, including human awareness of the importance of these trends in the environment. However, despite significant advances in the field, major obstacles remain and it is still very difficult to provide a relevant representation of the biodiversity at a global scale (Colwell, 2009). The study of marine microorganisms is no exception and has major difficulty related to the characteristics of the individuals being studied. Moreover, the notion of species along with size, variety, genetic proximity and more importantly ubiquity (Fenchel and Finlay, 2014) among these microbial organism demand an end-to-end systematic study. Considering the dichotomy between the dominant and rare types, it is possible to describe the characteristics of an environment with the shape of the rank abundance distribution and thereby this can provide a new biodiversity study tool. Studies have suggested a correlation between the chaotic movements and planktonic communities (Hernandez-Garcia and Lopez, 2004), which demonstrate a tendency towards power-law behavior (Stumpf and Porter, 2012). Using the plethora of sequencing data generated by Tara Oceans, the existence of a single slope suggests that common organizing processes may shape this seemingly universal feature of marine ecosystems. The identification of the mechanism that underpin such property in the world oceans is beyond the scope of this study. However, the preliminary results presented in this chapter have nonetheless set the stage to gain a fine insight into some key ecological questions (using RADs), for instance, does species rank influence contribution to functional diversity? How do rare species contribute to functional diversity? Can we decouple ecological and evolutionary processes for microbial species in an environment that undergoes constant and massive perturbations?

CHAPTER 6

General Conclusions and Future Perspectives

"Ecological patterns, about which we construct theories, are only interesting if they are repeated. They may be repeated in space or in time, and they may be repeated from species to species. A pattern which has all of these kinds of repetition is of special interest because of its generality, and yet these very general events are only seen by ecologists with rather blurred vision. The very sharp-sighted always find discrepancies and are able to say that there is no generality, only a spectrum of special cases. This diversity of outlook has proved useful in every science, but it is nowhere more marked than in ecology."

–Robert MacArthur, 1968

This study set out to explore global patterns of biodiversity of marine planktonic diatoms and to gain insights into the mechanisms involved in their community structure and assembly. The study also sought to examine Rank Abundance Distributions (RADs) within planktonic protistan communities to explore if there exists a universal power-law tail. Despite recent progress, our knowledge of the factors that account for biogeographic patterns remains limited. Owing to the enormous diversity prevalent in diatoms (and other microbial eukaryotes) in their natural environments, one of the major challenges remains the difficulty of fully addressing diversity, even with the advances in high-throughput sequencing technologies. This, in turn, limits our knowledge of the fundamental principles involved in microbial geography. However, diatoms are present ubiquitously and hence have allowed me to test microbial biogeographic theories to uncover the processes that might explain their diversity and distribution. To achieve an insight into biogeographical patterns, the key precept applied was to use a standardized DNA region, which can be used to identify known species and to aid in the discovery of undescribed ones (Hebert et al., 2003). The study sought to address the following key questions: to determine how diatom abundance and diversity vary in the world ocean, to evaluate if the diversity is consistent across different size classes and oceanic provinces, and to develop an understanding of how their communities are structured.

A metabarcoding approach for diversity assessment was developed (**Chapter 2**) in this thesis. Comparable overlap between the classical morphological method using light microscopy (LM) and molecular identification method illustrated that the metabarcoding approach offers a promising way to perform diatom diversity assessments. In some samples, it was found that a few genera could be identified only by light microscopy. On the other hand, it was also observed that a few genera were identified based on their sequence, not by their morphology. The former observation is probably due to the lack of representative sequences in the reference database, whereas the latter might occur either due to those genera being cryptic species or them being over-shadowed by a few excessively dominant genera. A more complete reference database will be of immense help in concluding the inferences here.

The unprecedented *Tara* Oceans dataset allowed a detailed evaluation of global diatom distribution and diversity. The saturating rarefaction curve indicated the completeness of the data. Further, the number of unobserved ribotypes evaluated by parametric estimators revealed that the sampling was 65 % complete. All such diversity analyses are always based on three assumptions. First, a sample is a representative of the whole community; second, the genetic marker diversity is representative of the overall organism diversity; third, when comparing samples from different environments, one assumes that all the biases, i.e., sampling, sequencing and diversity calculations, are similar across environments.

This first in-depth global study of diatoms using metabarcoding revealed that,

- diatom abundances and diversities show complex patterns,
- there is considerable unknown diversity within diatom communities,
- highest abundances were in Southern Ocean, Malvinas confluence, Benguela and Peruvian upwellings,
- diatom communities show ecological and biogeographical patterns,
- physical forcing is a major driver of diatom biodiversity,
- there is exceptionally high diatom diversity in the open ocean.

In addition, the distribution of well-studied genera was found to be consistent with other previously reported studies. The worldwide distribution of different ribotypes from the most abundant diatom genera is consistent with the fact that diatoms have evolved to adapt to varying environmental conditions to exploit a range of ecological niches.

The diatoms are known to favor nutrient rich coastal environments. In contrast to coastal boundaries, the open oceans are areas away from the coast and continental shelves and are highly heterogeneous and dynamic in nature. The remarkably high diatom diversity recorded in such an area in this study is surprising as such areas are deprived of the essential nutrients that are required for species sustenance and growth. Although diatoms are known to be a widespread group and can adapt to varying environmental conditions, it is expected that only a few well-adapted species could tolerate such conditions. Our study revealed high diversity in these zones, which in turn also suggests a larger number of species interactions which can further fundamentally affect ecosystem properties. To-date, our knowledge regarding diversity patterns of diatoms in the open ocean is very scarce.

The various physico-chemical parameters and other contextual data collected during the *Tara* Oceans expedition offered to develop an understanding of the processes that are involved in structuring marine diatom communities and controlling their biodiversity. Over the past decade, there has been an enduring interest in identifying simple rules and laws which define the observed complex phenomena in nature, for instance, the spatial distribution of a microbial organism in the ocean. This complexity of a dynamic phenomenon is a result of interplay of relationships among the units of communities, acting at a spatial and temporal scale. In the recent past, ecologists have proposed several restricted theories or empirical models, e.g., niche theories, dispersal theories, power-laws, to capture an insight into the underlying complex dynamic phenomenon (e.g., Hubbell, 2011; Hutchinson, 1957; Chase and Leibold, 2003; Levin et al., 2003). There has been a long on-going debate on whether these restricted theories alone can explain all the aspects, considering these processes being inter-dependent and non-linear. Collectively, these studies and the present thesis emphasize that the development and maintenance of ecological communities are controlled by multiple processes (e.g., **Figure 6.1)** that act together in an interactive manner. Here, I show that environmental heterogeneity is not always the only factor in structuring diatom communities but that dispersal limitation also mediates its structure (**Chapter 3**). The results have demonstrated that diatoms are not a randomly distributed entity at a large spatial scale, but rather represent a biogeographically structured ecological community, regulated by both environmental heterogeneity and spatial processes.

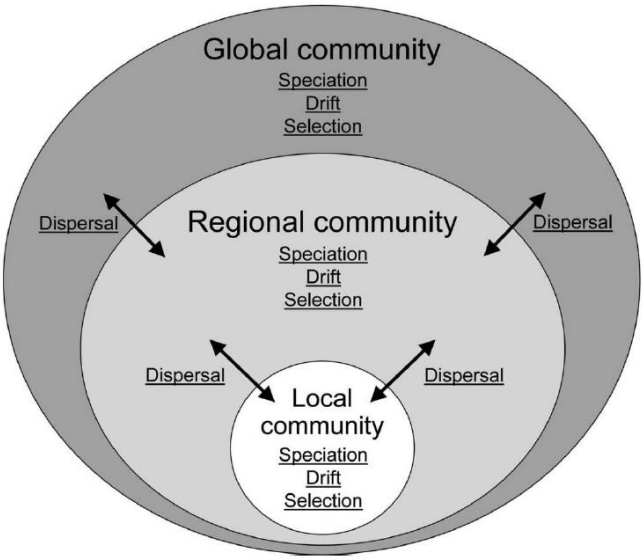
Next, the distinct behavior and response to environmental conditions were evaluated by expressing co-occurring ribotypes in significantly associated clusters. Each identified cluster was expressed as a function of varying environmental parameters (**Chapter 4**).

Finally, a broader study on the whole marine eukaryotic microbial community revealed that all the sampled communities followed comparable structural patterns. These were characterized by a few dominant ribotypes representing the majority of abundance and a large number of rare ribotypes representing a long tail (**Chapter 5**). Preliminary analyses demonstrated that the tail of the rank abundance distributions (RADs) exhibit a power-law behavior. However, detecting power-laws in these systems is subjected to caution as they are associated with chaotic events. Previous studies have emphasized the correlation between chaotic movements and planktonic communities which demonstrated a tendency towards power-law behavior.

One of the most enduring principles of microbial ecology over the years has been “everything is everywhere, but the environment selects” (Baas Becking, 1934). It reflects the notion that microbial species are not limited by dispersion, and that a site’s species profile results from winnowing down a

Conceptual construct I

(Source: Figure 4 from Vellend 2010).



Conceptual construct II

(Figure modified from HilleRisLambers et al. (2012))

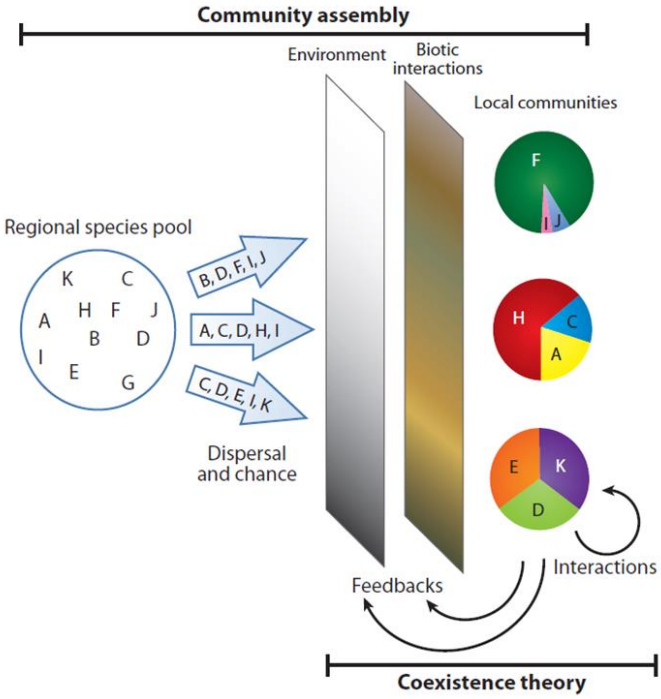


Figure 6.1. Examples of existing organizational frameworks in Community Ecology. ***Upper panel*****:** Selection, drift, speciation, and dispersal interact to determine community dynamics across spatial scales. Note that ecological drift as a key processes is important because it suggests that some patterns of abundance are simply stochastic which results from the demographic equivalence or similarity between species. ***Lower panel*****:** Typical filter model of community assembly.

comprehensive list of organisms based on environmental parameters. In this thesis, large-scale diversity gradients were examined using 46 sampling stations across the world ocean and the overlap between diatom community structures was examined at local and regional scales. According to Bass Becking, organisms well suited to a demanding environment would have the same universal access to it regardless of its distance from a similar, already-colonized site. On the contrary, this study showed that community dissimilarity increases with the increasing spatial distance between sampling stations, suggesting that there exists something that inhibits global dispersal forces. This finding is in line with a similar study reported for bacteria (Sul et al., 2013). A latitudinal diversity gradient was also found for diatom communities, as reported for bacteria by Sul et al. (2013). A more complete picture in this regard can be obtained in the future as the data from 170 distinct *Tara* Oceans stations spread globally across the world ocean will shortly be available. Nevertheless, this study demonstrated that in addition to environmental factors, ocean currents, turbulence and/or advection play key roles in governing diatom diversity. The revised proclamation (Sul et al., 2013): "*everything is not everywhere, and the environment selects*" appears appropriate also for marine diatom communities.

Various scientific reports have emphasized the key role of microbes in regulating Earth's climate. However, microbial diversity, in itself, is being altered by the ongoing human-induced climate change. As a result of this alteration, dominant species may become extinct and completely unknown species may become dominant. In recent years, there has been an increasing interest in the discovery of marine microbial diversity, but still there is a long way to go. The biggest challenge posed for studying them is that the vast majority (>90%) of these microbes cannot be cultured and hence cannot be studied with classical methods. Therefore, in the light of loss or degradation of marine biodiversity and impacts of climate change which rapidly alter ecosystems worldwide, there is an urgent need for standardized and comparable data in order to detect changes of biodiversity. The metabarcoding approach described in this thesis provides a method for a comprehensive diversity assessment and evaluation. This method has been shown to be representative as well as pragmatic.

Answers to various ecological questions and our increasing concerns about climate change and other environmental problems are major incentives for pursuing global marine diversity assessments and to unveil how these changes are altering ecosystems and their services. In years to come, decoding the ecological and evolutionary rules governing this exceptional diversity will be essential for understanding one the most critical biomes for the functioning of the Earth system. Under the largely human induced environmental change, development of predictive habitat distribution models will enable us to forecast how ecological systems will behave in the future. The impetus for this effort may

promote an understanding of the effects of environmental change on the world and what might be performed to mitigate or adapt to them.

Future research directions

Despite the limitations (e.g. seasonality, reference DB), this study has demonstrated not only the validity of the metabarcoding approach for diatoms, but also the ways in which diatoms are distributed in the world's oceans. I contend that my results demonstrate that beta diversity can reveal important aspects on a global scale, much in the same way that species richness is considered at either the alpha (sample) level or the gamma (regional) level (Chapter 2). Given that environmental heterogeneity demonstrated varying and significant impacts on diatoms of different size classes (Chapter 3), this study emphasizes to study not only species richness/composition, but how local communities vary across oceanic provinces.

Perhaps one of the most straightforward and immediately beneficial approaches based on this work will be to apply the framework developed herein to the forthcoming data sets from all the *Tara* Oceans sampling stations, that cover eight oceanic provinces, and to study distribution and diversity of other taxonomic lineages (de Vargas, 2015). Exploiting these data sets to gain a complete end-to-end diversity assessment increases the return on investments of both time and finances. I encourage other groups to test the methods herein on their existing datasets to get a better picture of the generality of the results I obtained.

In order to best describe the influences of niche-based environmental heterogeneity and spatial processes on community structure, it will be important to explore functional and phylogenetic diversity. Importantly, thanks to the continuing advances in computing power together with the availability of modifiable, open source computer codes and diversity tools, this is no longer an intimidating task. This synthesis will generate an even better picture of the processes that lead to the incredible patterns of organismal diversity observed around the world.

Furthermore, it will be important to validate the methods developed in Chapter 5. Clearly I have not fully tested the metrics. In addition, to improve upon the work conducted herein, the following recommendations can be made for future studies:

- Known unknowns. To explore the identity of these novel ribotypes, cloning and sequencing larger portions of the corresponding rDNA gene is recommended.

- Building an exhaustive and representative reference databases represents the most critical issue limiting sequence assignation. In the future, continued efforts in this direction will lead us towards a more complete quantification of novelty and diversity.
- Disentangling the effect of highly correlated environmental variables by independently modelling associations of each of the two predictor variables with taxonomic/functional composition is recommended. This can be achieved using the approach described by Sunagawa et al. (2015).
- Investigating the influence of within- and cross-kingdom biological interactions on community heterogeneity.
- Considering the dichotomy between the dominant and rare types, describing the characteristics of an environment with the shape of the rank abundance distribution.
- Investigating the impact of clustering on the structure of the tail of RAD which may facilitate an understanding of the ongoing evolutionary processes at the molecular level.

With the availability of an unprecedented *Tara* oceans (2009-2013) metabarcoding dataset from 210 stations covering eight major oceanic provinces, the predictive habitat distribution models remains another interesting perspective for future work because diversity studies are pivotal in providing insights regarding how their richness and community composition contribute to ecosystem function. Such studies may support the development of predictive habitat distribution modeling that describe how microbial communities will respond/change to natural or anthropogenically mediated changes in environmental conditions (Caron et al., 2012). These models are generally based on various hypotheses as to how environmental factors control the distribution of species and communities. There is therefore ample opportunity to expand on the field of biogeography.

"The history of the planet,
and so the humanity,
lies in the heart of the oceans.
And its future too."

References

Adl MS, Leander BS, Simpson AJM, Anderson OR, Barta JR, Bass D, Bowser SS, Brugerolle G, Farmer MA, Karpov S, Kolisko M, Lane CE, Lodge J, Lynn DH, Mann DG, Meisterfeld R, Mendoza L, Moestrup Ø, Mozley-Standridge SE, Smirnov AV, Spiegel FW (2007) Diversity, Nomenclature and Taxonomy of Protists. Syst Biol 56(4):684-689.

Adler RJ, Feldman RE, Taqqu MS, Eds. (1998) A Practical Guide to Heavy Tails: Statistical Techniques and Applications. (Birkhauser, Boston).

Allen AE, Smith CL, Oborník M, Horák A, Nunes-Nesi A, McCrow JP, Zheng H, Johnson DA, Hu H, Fernie AR, Bowler C (2011) Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. Nature 473:203-207.

Amaral-Zettler LA, McCliment E, Huse S (2009) A method for studying protistan diversity using massively parallel sequencing of V9 hyper variable regions of small-subunit ribosomal RNA genes. PLoS one 4:e6372.

Amaral-Zettler LA, Zettler ER, Theroux SM, Palacios C, Aguilera A, Amils R (2011) Microbial community structure across the tree of life in the extreme Rio Tinto. ISMEJ 5:42-50.

Amarasekare P, Nisbet R (2001) Spatial heterogeneity, source-sink dynamics and the local coexistence of competing species. Am Nat 158:572-584.

Anderson MJ, Crist TO, Chase JM, Vellend M, Inouye BD, Freestone AL, Sanders NJ, Cornell HV, LS Comita, Davies KF, Harrison SP, Kraft NJB, Stegen JC, Swenson NG (2011) Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. Ecol Lett 14:19-28.

Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, et al. (2006) The Marine Viromes of Four Oceanic Regions. PLoS Biol 4(11):e368.

Anonymous (1703) Two letters from a gentleman in the country, relating to Mr. Leuwenhoeck's letter in Transaction, No. 283. Philos Trans R Soc Lond 23 (288):1494.

Appeltans W, Ahyong ST, Anderson G, Angel MV, Artois T, Bailly N, Bamber R, Barber A, Bartsch I, Berta A, et al. (2012) The Magnitude of Global Marine Species Diversity. Curr Biol 22:2189-2202.

Appeltans W, Ahyong ST, Anderson G, Angel MV, et al. (2012) The magnitude of global marine species diversity. Curr Biol 22:2189-202.

Armbrust EV (2009) The life of diatoms in the world's oceans. Nature 459:185-192.

Armbrust EV, Berges JA, C Bowler, Green BR, et al. (2004) The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. Science 306:79-86.

Baas Becking LGM (1934) Geobiologie of inleiding tot de milieukunde. The Hague, the Netherlands: W.P. Van Stockum & Zoon (in Dutch).

Bahram M, Koljalg U, Courty PE, Diedhiou AG, Kjøller R, Polme S, Ryberg M, Veldre V, Tedersoo L (2013) The distance decay of similarity in communities of ectomycorrhizal fungi in different ecosystems and Scales. J Ecol 101:1335-1344.

Bahram M, Põlme S, Kõljalg U, Zarre S, Tedersoo L (2012) Regional and local patterns of ectomycorrhizal fungal diversity and community structure along an altitudinal gradient in the

Hyrcanian forests of northern Iran. New Phytol 193:465-473.

Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality: An explanation of the 1/f noise. Phys Rev Lett 59:381-384.

Balvanera P, Pfisterer AB, Buchmann N, He JS, Nakashizuka T, Raffaelli D et al. (2006) Quantifying the evidence for biodiversity effects on ecosystem functioning and services. Ecol Lett 9:1146-1156.

Barange M, Campos B (1991) Models of species abundance- a critique and an alternative to the dynamics model. Mar Ecol Prog Ser 69:293-298.

Barton AD, Dutkiewicz S, Flierl G, Bragg J, Follows MJ (2010) Patterns of diversity in marine phytoplankton. Science 327:1509-1511.

Bauke H (2007) Parameter estimation for power-law tail distributions by maximum likelihood methods. Eur Phys J B 58:167-173.

Begossi A (1996) Use of ecological methods in ethnobotany: Diversity indices. Econ Bot 50(3):280-289.

Behnke A, Engel M, Christen R, Nebel M, Klein RR, Stoeck T (2011) Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. Environ Microbiol 13: 340-349.

Behnke A, Friedl T, Chepurnov VA, Mann DG (2004) Reproductive compatibility and rDNA sequence analyses in the *Sellaphora pupula* species complex (Bacillariophyceae). J Phycol 40:193-208.

Bellemain E, Carlsen T, Brochmann C, Coissac E, Taberlet P, Kauserud H (2010) ITS as an environmental DNA barcode for fungi: an *in silico* approach reveals potential PCR biases. BMC Microbiol 10:189.

Beszteri B, John U, Medlin LK (2007) An assessment of cryptic genetic diversity within the *Cyclotella meneghiniana* species complex (Bacillariophyta) based on nuclear and plastid genes, and amplified fragment length polymorphisms. Ann Eur J Phyco 42:47-60.

Bik HM, Halanych KM, Sharma J, Thomas WK (2012) Dramatic shifts in benthic microbial eukaryote communities following the deepwater horizon oil spill. PLoS one 7(6): e38550.

Birks HJB (2010) Numerical methods for the analysis of diatom assemblage data. The diatoms: applications for environmental and earth sciences, eds Smol JP, Stoermer EF (Cambridge University press, Cambridge), pp 23–54.

Bittner L, Gobet A, Audic S, Romac S, Egge ES, Santini S, Ogata H, Probert I, Edvardsen B, de Vargas C (2013) Diversity patterns of uncultured Haptophytes unravelled by pyrosequencing in Naples Bay. Mol Ecol 22:87-101.

Blanchet FG, Legendre P, Borcard D (2008) Forward selection of explanatory variables. Ecology 89(9):2623-2632.

Blanco S, Ector L, Bécares E (2004) Epiphytic diatoms as water quality indicators in Spanish shallow lakes. Vie et Milieu 54:71–79.

Blois JL, Zarnetske PL, Fitzpatrick MC, Finnegan S (2013) Climate change and the past, present, and future of biotic interactions. Science 341(6145):499-504.

Boltovskoy D, Correa N, Boltovskoy A (2005) Diversity and endemism in cold waters of the South Atlantic: contrasting patterns in the plankton and the benthos. Sci Mar 69:17-26.

Bopp L, Aumont O, Cadule P, Alvain S, Gehlen M (2005) Response of diatoms distribution to global warming and potential implications: A global model study. Geophys Res Lett 32:1-4.

Bouman HA, Ulloa O, Scanlan DJ, Zwirglmaier K et al. (2006) Oceanographic basis of the global surface distribution of Prochlorococcus ecotypes. Science 312:918-921.

Bowler C, et al. (2008) The *Phaeodactylum* reveals the evolutionary history of diatom genomes. Nature 456:239-244.

Bragg JG, Dutkiewicz S, Jahn O, Follows MJ, Chisholm SW (2010) Modeling selective pressures on phytoplankton in the global ocean. PLoS one 5:e9569.

Bremnera J, Rogersb SI, Frida CLJ (2006) Methods for describing ecological functioning of marine benthic assemblages using biological traits analysis (BTA). Ecol Indic 6(3):609–622

Brown JH, West GB (2000) Scaling in Biology. (Oxford University Press, New York).

Brown MV, Ostrowski M, Grzymski JJ, Lauro FM (2014) A trait based perspective on the biogeography of common and abundant marine bacterioplankton clades. Mar Genomics 15:17-28.

Brussaard CPD, Payet JP, Winter C, Weinbauer MG (2010) Quantification of aquatic viruses by flow cytometry. Manual of Aquatic Viral Ecology, eds Wilhelm SW, Weinbauer MG, Suttle CA (ASLO), pp 102-109.

Bucklin A, Steinke D, Blanco-Bercial L (2011) DNA Barcoding of Marine Metazoa. Ann Rev Marine Sci 3:471-508.

Buerki S, Forest F, Stadler T, Alvarez N (2013) The abrupt climate change at the Eocene-Oligocene boundary and the emergence of South-East Asia triggered the spread of sapindaceous lineages. Ann Bot 112(1):151-160.

Buitenhuis ET, Li WKW, Lomas MW, Karl DM, Landry MR, Jacquet S (2012) Picoheterotroph (*Bacteria* and *Archaea*) biomass distribution in the global ocean. Earth Syst Sci Data 4:101-106.

Burki F, Keeling PJ (2014) Rhizaria. Curr Biol 24:R103-107.

Butcher RW (1947). Studies in the ecology of rivers. IV. The algae of organically enriched water. J Ecol 35:186–91.

Cardinale BJ, Srivastava DS, Duffy JE, Wright JP, Downing AL, Sankaran M, et al. (2006) Effects of biodiversity on the functioning of trophic groups and ecosystems. Nature 443:989-992.

Cardinale BJ, Wright JP, Cadotte MW, Carroll IT, Hector A, Srivastava DS, et al. (2007) Impacts of plant diversity on biomass production increase through time because of species complementarity. Proc Natl Acad Sci USA 104:18123-18128.

Carlson JM, Doyle J (1999) Highly optimized tolerance: a mechanism for power laws in designed systems. Phys Rev E 60:1412-1427.

Cavalier-Smith T (1999) Principles of protein and lipid targeting in secondary symbiogenesis: eugleno-

-id, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. J Eukaryot Microbiol 46:347-366.

Cermeño P, de Vargas C, Abrantes FT, Falkowski PG (2010) Phytoplankton biogeography and community stability in the ocean. PLoS one 5:e10037.

Cermeño P, Falkowski PG (2009) Controls on diatom biogeography in the ocean. Science 325:1539-1541.

Cervato C, Burckle L (2003) Pattern of first and last appearance in diatoms: Oceanic circulation and the position of polar fronts during the Cenozoic. Paleoceanography 18:1055.

Chamnansinp A, Li Y, Lundholm N, Moestrup Ø (2013) Global diversity of two widespread, colony-forming diatoms of the marine plankton, *Chaetoceros socialis* (syn. C. radians) and *Chaetoceros gelidus* sp. nov. J Phycol 49:1128-1141.

Chapin FS III, Zaveleta ES, Eviner VT, Naylor RL, Vitousek PM, Lavorel S, Reynolds HL, Hooper DU, Sala OE, Hobbie SE, Mack MC, Diaz S (2000) Consequences of changing biotic diversity. Nature 405:234-242.

Charles DF (1985) Relationships between surface sediment diatom assemblages and lake water characteristics in Adirondack lakes. Ecology 66(3):994-1011.

Chase JM (2005) Towards a really unified theory for metacommunities. Funct Ecol 19:182-186.

Chave J (2004) Neutral theory and community ecology. Ecol Lett 7:241-253.

Chave J, Leigh EG (2002) A spatially explicit neutral model of beta-diversity in tropical forests. Theor Popul Biol 62:153-168.

Chesson P (2000) Mechanisms of maintenance of species diversity. Annu Rev Ecol Syst 31: 343-366.

Church MJ (2008) Resource control of bacterial dynamics in the sea. Microbial Ecology of the Oceans, Second Edition, ed Kirchman DL (John Wiley & sons Inc., New York), pp 335-382.

Chust G, Irigoien X, Chave J, Harris RP (2013) Latitudinal phytoplankton distribution and the neutral theory of biodiversity. Global Ecol Biogeogr 22(5):531-543.

Clarke A, Crame JA (1997) Diversity, latitude and time: patterns in the shallow sea. Marine Biodiversity, eds Ormond RFG, Gage JD, Angel MV (Cambridge University Press, Cambridge). pp 122–147.

Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. SIAM Review 51(4):661-703.

Cloern JE, Dufford R (2005) Phytoplankton community ecology: principles applied in San Francisco Bay. Mar Ecol Prog Ser 285:11-28.

Coissac E, Riaz T, Puillandre N (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. Mol Ecol 21(8):1834-1847.

Colling A (2001) Ocean Circulation, Open University Course Team. Second Edition. ISBN 978-0-7506-5278-0.

Colwell RK (2009) Biodiversity: concepts, patterns and measurement. Princeton Guide to Ecology, ed Levin SA (Princeton University Press, Princeton, NJ) pp 257-263.

Cornell HV (1999) Unsaturation and regional influences on species richness in ecological communities: a review of the evidence. Ecoscience 6:303-315.

Cottenie K (2005) Integrating environmental and spatial processes in ecological community dynamics. Ecol Lett 8:1175-1182.

Countway PD, Caron DA (2006) Abundance and distribution of *Ostreococcus* sp. in the San Pedro Channel, California, as revealed by quantitative PCR. Appl Environ Microbiol 72:2496.

Crosta X, Romero O, Armand LK, Pichon JJ (2005) The biogeography of major diatom taxa in Southern Ocean sediments: 2. Open ocean related species. Palaeogeogr Palaeoclimatol Palaeoecol 223:66-92.

Cullen JJ, Franks PJS, Karl DM, Longhurst AR (2002) Physical influences on marine ecosystem dynamics. Biological-physical interactions in the sea, eds Robinson AR, McCarthy JJ, Rothschild BJ (Wiley, New York), pp297-336.

Cunningham SA, Alderson SG, King BA (2003) Transport and variability of the Antarctic Circumpolar Current in Drake Passage. J Geophys Res 108:8084.

d'Ovidio F, De Monte S, Alvain S, Dandoneau Y, Lévy M (2010) Fluid dynamical niches of phytoplankton types. Proc Natl Acad Sci USA 107 (43):18366.

Dale B (1977) Cysts of toxic red tide dinoflagellate Conyaulax excavata (Braarud) Belech from Oslofjorden, Norway. Sarsia 63:29-34.

Dayton PK, Morbida BJ, Bacon F (1994) Polar marine communities. Amer Zool:3490-3499.

De Bie T, De Meester L, Brendonck L, Martens K, Goddeeris B, et al. (2012) Body size and dispersal mode as key traits determining metacommunity structure of aquatic organisms. Ecol Lett 15:740-747.

de Vargas C, Audic S, Henry N, Decelle J, et al. (2015) Eukaryotic plankton diversity in the sunlit global ocean. Science (In press).

de Vargas C, Norris R, Zaninetti L, Gibb SW, Pawlowski J (1999) Molecular evidence of cryptic speciation in planktonic foraminifers and their relation to oceanic provinces, Proc Natl Acad Sci USA 96:2864-2868.

Degerlund M, Huseby S, Zingone A, Sarno D, Landfald B (2012) Functional diversity in cryptic species of *Chaetoceros socialis Lauder* (Bacillariophyceae). J Plank Res 34: 416-431.

Delwiche CF (1999) Tracing the thread of plastid diversity through the tapestry of life. Am Nat 154:S164-S177.

Denman KL (2008) Climate change, ocean processes and ocean iron fertilization. Mar Ecol Prog Ser 364:219-225.

d'Ovidio F, De Monte S, Alvain S, Dandonneau Y, Lévy M (2010) Fluid dynamical niches of phytoplankton types. Proc Natl Acad Sci USA 107:18366-18370.

Drake JA (1990) The mechanics of community assembly and succession. J Theor Biol 147:213-233.

Dray S, Legendre P, Peres-Neto P (2006) Spatial modeling: a comprehensive framework for principal coordinate analysis of neighbor matrices (PCNM). Ecol Model 196:483-493.

Ducklow HW (2000) Bacterioplankton production and biomass in the oceans. Microbial Ecology of the Oceans, ed Kirchman DL (Wiley,New York), pp 85-120.

Durrett R, Levin S (1996) Spatial models for species area curves. J Theor Biol 179:119-127.

Edgcomb V, Orsi W, Bunge J, Jeon S, Christen R, Leslin C, et al. (2011a). Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. ISMEJ 5:1344-1356.

Edlund MB, Stoermer EF (1997) Ecological, evolutionary, and systematic significance of diatom life histories. J Phycol 33:897-918.

Evans C, Archer SD, Jacquet S, Wilson WH (2003) Direct estimates of the contribution of viral lysis and microzooplankton grazing to the decline of a *Micromonas* spp. population. Aquat Microb Ecol 30:207-219.

Evans KM, Hayes PK (2004) Microsatellite markers for the cosmopolitan marine diatom *Pseudo-nitzschia pungens*. Mol Ecol Notes 4:125-126.

Evans KM, Kuhn SF, Hayes PK (2005) High levels of genetic diversity and low levels of genetic differentiation in North Sea *Pseudo-nitzschia pungens* (Bacillariophyceae) populations. J Phycol 41:506-514.

Evans KM, Mann DG (2009) A proposed protocol for nomenclaturally effective DNA barcoding of microalgae. Phycologia 48(1):70-74.

Evans KM, Wortley AH, Mann DG (2007) An assessment of potential diatom "barcode" genes (cox1, rbcL, 18S and ITS rDNA) and their effectiveness in determining relationships in *Sellaphora* (*Bacillariophyta*). Protist 158(3):349-364.

Falkowski PG (2002) The ocean's invisible forest - Marine phytoplankton play a critical role in regulating the earth's climate. Could they also be used to combat global warming. Sci Am 287(2):54-61.

Falkowski PG, Barber RT, Smetacek V (1998) Biogeochemical controls and feedbacks on ocean primary production. Science 281 (5374):200-206.

Falkowski PG, et al. (2005) The rise of oxygen over the past 205 million years and the evolution of large placental mammals. Science 309:2202-2204.

Falkowski PG, Knoll AH (Ed.) (2011) Evolution of Primary Producers in the Sea (Elsevier, Boston). ISBN: 978-0-12-370518-1.

Farjalla VF, Srivastava DS, Marino NAC, Azevedo FD, Dib V, et al. (2012) Ecological determinism increases with organism size. Ecology 93:1752-1759.

Fenchel T, Esteban GF, Finlay BJ (1997) Local versus global diversity of microorganisms: cryptic diversity of ciliated protozoa. Oikos 80:220-225.

Fenchel T, Finlay BJ (2004) The ubiquity of small species: patterns of local and global Diversity. Bioscience 54:777-784.

Fernandes LF, Calixto-Feres M (2012) Morphology and distribution of two epizoic diatoms (Bacillariophyta) in Brazil. Acta Bot Bras 26(4):836-841.

Ferriere R, Cazelles B (1999) Universal power laws govern intermittent rarity in communities of interacting species. Ecology 80:1505-1521.

Ficetola GF, Coissac E, Zundel S, Riaz T, Shehzad W, Bessiere J, Taberlet P, Pompanon F (2010) An *in silico* approach for the evaluation of DNA barcodes. BMC Genom 11:434.

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P (1998) Primary production of the biosphere: Integrating terrestrial and oceanic components. Science 281:237-240.

Finlay BJ (2002) Global dispersal of free-living microbial eukaryote species. Science 296:1061-1063.

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

Finlay BJ, Fenchel T (1999) Divergent perspectives on protist species richness. Protist 150:229-233.

Fisher CK, Mehta P (2014) The transition between the niche and neutral regimes in ecology. Proc Natl Acad Sci USA 111(36):13111-13116.

Fjerdingstad E (1950) The microflora of the River Molleaa with special reference to the relation of benthic algae to pollution. Folia Limnologica Scandanavica 5:1–123.

Fourtanier E, Kociolek JP (2003) Addendum to the catalogue of diatom generic names. Diatom Res 18:245-258.

Fuhrman JA (2009) Microbial community structure and its functional implications. Nature 459:193-199.

Fuhrman JA, Steele JA, Hewson I, Schwalbach MS, Brown MV, Green JL, Brown JH (2008) A latitudinal diversity gradient in planktonic marine bacteria. Proc Natl Acad Sci USA 105(22):7774-7778.

Gallagher JC (1980) Population genetics of *Skeletonema costatum* (*Bacillariophyceae*) in Narragansett bay. J Phycol 16:464-474.

Gallienne CP, Robins DB (2001) Is Oithona the most important copepod in the world's oceans? J Plankton Res 23(12):1421-1432.

Galluzzi L, Penna A, Bertozzini E, Vila M, Garcés E, Magnani M (2004) Development of a real-time PCR assay for rapid detection and quantification of *Alexandrium minutum* (a dinoflagellate). Appl Environ Microbiol 70:1199-1206.

Garate-Lizarraga I, Muneton-Gomez MDS (2009) Primer registro de la diatomea epibionte *Pseudohimantidium pacificum* y de otras asociaciones simbióticas en el Golfo de California. Act Bot Mex (88):31-45.

Gaston KJ (1996) Biodiversity - latitudinal gradients. Prog Phys Geogr 20:466-476.

Gaston KJ (Ed.) (1994) Rarity. (Chapman & Hall, London).

Gaston KJ, Chown SL (2005) Neutrality and the niche. Funct Ecol 19:1-6.

Gelashvili DB, Iudin DI, Rozenberg GS, Iakimov NV (2007) Power-law species richness accumulation as manifestation of the fractal community structure. Zh Obshch Biol 68(3):170-9.

Georgianna DR, Mayfield SP (2012) Exploiting diversity and synthetic biology for the production of algal biofuels. Nature 488(7411):329-335.

Gerbi SA (Ed.) (1985) Evolution of ribosomal DNA. Molecular Evolutionary Genetics (Plenum, New York, USA), pp 419-517.

Gersonde R, Zielinski U (2000) The reconstruction of late quaternary antarctic sea-ice distribution - the use of diatoms as a proxy for sea-ice. Palaeogeogr Palaeoclimatol Palaeoecol 162(34):263-286.

Gianoulis TA, Raes J, Patel PV, Bjornson R, Korbel JO, et al. (2009) Quantifying environmental adaptation of metabolic pathways in metagenomics. Proc Natl Acad Sci USA 106:1374-1379.

Gilabert J (2001) Short-term variability of the planktonic size structure in a Mediterranean coastal lagoon. J Plankton Res 23:219-226.

Giovannoni SJ, Stingl U (2005) Molecular diversity and ecology of microbial plankton. Nature 437:343-348.

Godhe A, Asplund ME, Harnstrom K, Saravanan V, Tyagi A, Karunasagar I (2008) Quantifying diatom and dinoflagellate biomass in coastal marine sea water samples by real-time PCR. Appl Environ Microbiol 74:7174-7182.

Godhe A, McQuoid MR, Karunasagar I, Karunasagar I, Rehnstam-Holm AS (2006) Comparison of three common molecular tools for distinguishing among geographically separated clones of the diatom *Skeletonema marinoi Sarno et Zingone* (*Bacillariophyceae*). J Phycol 42:280-291.

Goldstein ML, Morris SA, Yen GG (2004) Problems with fitting to the power-law distribution. Eur Phys J B 41:255-258.

Göthe E, Angeler DG, Gottschalk S, Löfgren S, Sandin L (2013) The Influence of Environmental, Biotic and Spatial Factors on Diatom Metacommunity Structure in Swedish Headwater Streams. PLoS one 8(8):e72237.

Goto DK, Yan T (2011) Genotypic diversity of Escherichia coli in the water and soil of tropical watersheds in Hawaii. Appl Environ Microbiol 77(12):3988-3997.

Gravel D, Canham CD, Beaudet M, Messier C (2006) Reconciling niche and neutrality: the continuum hypothesis. Ecol Lett 9:399-409.

Griffiths D (1997) Local and regional species richness in North American lacustrine fish. J Anim Ecol 66:49-56.

Grimm V, Reise K, Strasser M (2003) Marine metapopulations: a useful concept? Helgol Mar Res 56:222–228.

Grömping U (2006) Relative Importance for Linear Regression in R: The Package *relaimpo*. J Stat Softw 17(1):1-27.

Guidry MW, Arvidson RS, MacKenzie FT (2007) Biological and geochemical forcings to Phanerozoic change in seawater, atmosphere, and carbonate precipitate composition. Evolution of Primary Producers in the Sea, eds Falkowski PG & Knoll AH (Elsevier, Boston), pp 377-403.

Guillou L, Bachar D, Audic S, Bass D, Berney C, et al. (2013) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Res 41(D1):D597-D604.

Guiry MD (2012) How many species of algae are there? J Phycol 48:1057-1063.

Haegeman B, Hamelin J, Moriarty J, Neal P, Dushoff J, Weitz JS (2013) Robust estimation of microbial diversity in theory and in practice. ISMEJ 7:1092-1101.

Hajibabaei M, Shokralla S, Zhou X, Singer GAC, Baird DJ (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. PloS one 6(4):e17497.

Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA (2007) DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. Trends Genet 23(4):167-172.

Halbert KMK, Goetze E, Carlon DB (2013) High cryptic diversity across the global range of the migratory planktonic copepods *Pleuromamma piseki* and *P. gracilis*. PLoS one 8:e77011.

Hall IH, Smol AJ (1992) A Weighted-averaging regression and calibration model for inferring total phosphorus concentration from diatoms in British Columbia (Canada) lakes. Freshwater Biol 27:417-434.

Halme P, Ódor P, Christensen M, Piltaver A, Veerkamp M, Walleyn R, Siller I, Heilmann-Clausen J (2013) The effects of habitat degradation on metacommunity structure of wood-inhabiting fungi in European beech forests. Biol Cons 168:24-30.

Hannon M, Gimpel J, Tran M, Rasala B, Mayfield S (2010) Biofuels from algae: challenges and potential. Biofuels 1(5):763-784.

Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JBH (2012) Beyond biogeographic patterns: processes shaping the microbial landscape. Nat Rev Microbiol 10(7):497-506.

Hartmann M, Howes CG, Abarenkov K, Mohn WW, Nilsson RH (2010). V-Xtractor: An open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16S/18S) ribosomal RNA gene sequences. J Microbiol Meth 83(2):250-253.

Hasle GR, Syvertsen EE, Steidinger KA, Tangen K (1996) Marine Diatoms. Identifying Marine Diatoms and Dinoflagellates, ed Tomas CR (Academic Press, New York), pp 429.

Hebert PDN, Cywinska A, Ball SL, De Waard JR (2003) Biological identifications through DNA barcodes. Proc R Soc B 270:313-321.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London (B) 270:313-332. A biogeochemical study of the coccolithophore, Emiliania huxleyi, in the North Atlantic. Geophys Res Lett 7:879

Hebert PDN, Gregory TR (2005) The promise of DNA barcoding for taxonomy. Syst Biol 54(5):852-859.

Hedlund BP, Staley JT (2003) Microbial endemism and biogeography. Microbial Diversity and Bioprospecting, ed Bull AT (ASM, Washington DC).

Hernandez-Garcia E, Lopez C (2004) Sustained plankton blooms under open chaotic flows. Ecol Complex 1(3):253-259.

Hill BM (1975) A simple general approach to inference about the tail of a distribution. Ann Statist 3:1163-1174.

Hillebrand H, Matthiessen B (2009) Biodiversity in a complex world: consolidation and progress in functional biodiversity research. Ecol Lett 12:1405-1419.

HilleRisLambers J, Adler PB, WS Harpole, Levine JM, Mayfield MM (2012) Rethinking Community Assembly through the Lens of Coexistence Theory. Ann Rev Ecol Evol System 43:227-248.

Hinder SL, Hays GC, Edwards M, Roberts EC, Walne AW, Gravenor MB (2012) Changes in marine dinoflagellate and diatom abundance under climate change. Nature Clim Change 2:271-275.

Hoffmann KH, Rodriguez-Brito B, Breitbart M, Bangor D, Angly F, Felts B, Nulton J, Rohwer F, Salamon P (2007) Power law rank-abundance models for marine phage communities. FEMS Microbiol Lett 273:224-228.

Holligan PM, Fernández E, Aiken J, et al. (1993) A biogeochemical study of the coccolithophore, *Emiliania huxleyi*, in the North Atlantic. Geophys Res Lett 7:879-900.

Hooper DU, Chapin FS, Ewel JJ, Hector A, Inchausti P, Lavorel S, et al. (2005) Effects of biodiversity on ecosystem functioning: a consensus of current knowledge. Ecol Monogr 75:3-35.

Horner RA (2002) A taxonomic guide to some common marine phytoplankton (Biopress), pp 25–30.

Hubbell SP (2005) Neutral theory and the evolution of functional equivalence. Ecology 87(6):1387-1398.

Hubbell SP (Ed.) (2001) The Unified Neutral Theory of Biodiversity and Biogeography. (Princeton University Press, Princeton, New Jersey, USA).

Huber JA, Welch DBM, Morrison HG, Huse SM, Neal PR, Butterfield DA et al. (2007) Microbial population structures in the deep marine biosphere. Science 318:97-100.

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310-3214.

Huseby S, Degerlund M, Zingone A, Hansen E (2012) Metabolic fingerprinting reveals differences between northern and southern strains of the cryptic diatom *Chaetoceros socialis*. Eur J Phycol 47:480-489.

Hwang UW, Kim W (1999) General properties and phylogenetic utilities of nuclear ribosomal DNA and mitochondrial DNA commonly used in molecular systematics. Korean J Parasitol 37(4):215-228.

Hyun JH, Kim KH (2003) Bacterial abundance and production during the unique spring phytoplankton bloom in the central Yellow Sea. Mar Ecol-Prog Ser 252:77-88.

Irigoien X, Huisman J, Harris RP (2004) Global biodiversity patterns of marine phytoplankton and zoo-

-plankton. Nature 429(6994):863-867.

Janech M, Krell A, Mock T, Kang JS, Raymond J (2006) Ice-bindingproteins from sea ice diatoms (Bacillariophyceae). J Phycol 42(2):410-416.

Ji Y, Ashton L, Pedley SM, Edwards DP, Tang Y, Nakamura A, Kitching RL, Dolman P, Woodcock P, Edwards FA, Larsen TH, Hsu WW, Benedick S, Hamer KC, Wilcove DS, Bruce C, Wang XY, Levi T, Lott M, Emerson BC, Yu DW (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecol Lett 16(10):1245-1257.

John DW, Robert GS (2003) Freshwater algae of North America: ecology and classification. (Academic, San Diego California, USA), pp 917.

Jongman RHG, ter Braak CJF, van Tongeren OFR (1995) Data Analysis in Community and Landscape Ecology (Cambridge University Press, Cambridge).

Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C, et al. (2011) A holistic approach to marine eco-systems biology. PLoS Biology 9:e1001177.

Keeling PJ (2010) The endosymbiotic origin, diversification and fate of plastids. Philos Trans R Soc Lond B Biol Sci. 365(1541):729-748.

Ki JS, Han MS (2005) Molecular analysis of complete SSU to LSU rDNA sequence in the harmful dinoflagellate *Alexandrium tamarense* (Korean isolate, HY970328M). Ocean Sci J 40:155-166.

Knauss JA (2005) Introduction to Physical Oceanography, Second Edition (Waveland Press, Prospect Heights, IL).

Kooistra WH, Sarno D, Balzano S, Gu H, Andersen RA, Zingone A (2008) Global diversity and biogeography of *Skeletonema* species (*Bacillariophyta*). Protist 159:177-193.

Kooistra WHCF, De Stefano M, Mann DG, Medlin LK (2003) The phylogeny of the diatoms. Prog Mol Subcell Biol 33:59-97.

Kooistra WHCF, Gersonde R, Medlin LK, Mann DG (2007) The origin and evolution of the diatoms: their adaptation to a planktonic existence. Evolution of Primary Producers in the Sea, eds Falkowski PG & Knoll AH (Elsevier Academic Press, Singapore), pp 207-49.

Lazier JRN (2006) Dynamics of Marine Ecosystems (Blackwell Publishing), pp 35.

Leblanc K, Arístegui J, Armand L, Assmy P, Beker B, Bode A, Breton E, Cornet V, Gibson J, Gosselin MP, Kopczynska E, Marshall H, Peloquin J, Piontkovski S, Poulton AJ, Quéguiner B, Schiebel R, Shipe R, Stefels J, van Leeuwe MA, Varela M, Widdicombe C, Yallop M (2012) A global diatom database - abundance, biovolume and biomass in the world ocean. Earth Syst Sci Data 4:149-165.

Legendre P (2000) Comparison of permutation methods for the partial correlation and partial Mantel tests. J Stat Comput Simul 67:37-73.

Legendre P (2005) Species associations: the Kendall coefficient of concordance revisited. J Agric Biol Environ Stat 10(2):226-245.

Legendre P, Legendre L (ed)(1998) Numerical Ecology, 2nd ed (Elsevier Science, Amsterdam).

Leibold MA, McPeek MA (2006) Coexistence of the niche and neutral perspectives in community ecology. Ecology 87:1399-1410.

Leibold MA, Norberg J (2004) Biodiversity in metacommunities: Plankton as complex adaptive systems? Limnol Oceanogr 49(4):1278-1289.

Li WKW (2002) Macroecological patterns of phytoplankton in the northwestern North Atlantic Ocean. Nature 419:154-157.

Li WKW (2007) Annual average abundance of heterotrophic bacteria and *Synechococcus* in surface ocean waters. Limnol Oceanogr 43:1746-1753.

Li X, Holland DM, Gerber EP, Yoo C (2014) Impacts of the north and tropical Atlantic Ocean on the Antarctic Peninsula and sea ice. Nature 505 (7484):538-542.

Lichstein J (2007) Multiple regression on distance matrices: A multivariate spatial analysis tool. Plant Ecol 188:117-131.

Lindström ES, Östman Ö (2011) The Importance of Dispersal for Bacterial Community Composition and Functioning. PLoS one 6(10):e25883.

Litchman E, Klausmaier CA, Shofield OM, Falkowski PG (2007) The role of functional traits and trade-offs in structuring phytoplankton communities: Scaling from cellular to ecosystem level. Ecol Lett 10:1170-1181.

Longhurst A, Sathyendranath S, Platt T, et al. (1995). An estimate of global primary production in the ocean from satellite radiometer data. J Plankton Res 17:1245-1271.

Longhurst AR (1995) Seasonal cycles of pelagic production and consumption. Prog Oceanogr 36:77-67.

Longhurst AR (2006) Ecological Geography of the Sea, 2nd Edition (Academic Press, San Diego) pp 560.

Loreau M, Naeem S, Inchausti P (2002) Biodiversity and Ecosystem Functioning: Synthesis and Perspectives (Oxford University Press).

Lundholm N, Bates SS, Baugh KA, Bill BD, Connell LB, Léger C, Trainer VL (2012) cryptic and pseudo-cryptic diversity in diatoms-with descriptions of *pseudo-nitzschia hasleana* sp. Nov. and *P. fryxelliana sp. Nov*. J Phycol 48:436-454.

Lundholm N, Moestrup O, Kotaki Y, Hoef-Emden K, Scholin C, Miller P (2006) Inter- and intraspecific variation of the *Pseudo-nitzschia delicatissima*-complex (Bacillariophyceae) illustrated by rRNA probes, morphological data and phylogenetic analyses. J Phycol 42:464-481.

MacArthur RH (1957) On the relative abundance of bird species. Proc Nat Acad Sci USA 43:293-295.

Magurran AE (2004) Measuring Biological Diversity, (Blackwell Publishing).

Magurran AE, Henderson PA (2003) Explaining the excess of rare species in natural species abundance distributions. Nature 422:714-716.

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.

Maloney KO, Munguia P (2011) Distance decay of similarity in temperate aquatic communities: effects of environmental transition zones, distance measure, and life histories. Ecography 34:287-295.

Mann A (1917) The Economic Importance of the Diatoms. Volume 2465 of Publication (Smithsonian Institution). U.S. Government Printing Office, Harvard University. Pp 16.

Mann DG (1999). The species concept in diatoms. Phycologia 38:437-495.

Mann DG, Droop SJM (1996) Biodiversity, biogeography and conservation of diatoms. Hydrobiologia 336:19-32.

Mann DG, Sato S, Trobajo R, Vanormelingen P, Souffreau C (2010) DNA barcoding for species identification and discovery in diatoms. Cryptogamie Algologie 31(4):557-577.

Mann DG, Vanormelingen P (2013) An inordinate fondness? The number, distributions, and origins of diatom species. J Eukaryot Microbiol 60(4):414-420.

Mann KH, Lazier JR (2006) Dynamics of Marine Ecosystems, Third Edition (Blackwell Publishing).

Manneville P, Pomeau Y (1979) Intermittency and the Lorenz model. Phys Lett 75A 1.

Mao CX (2007) Estimating species accumulation curves and diversity indices. Statist Sinica 17:761-774.

Margalef R (1978) Life forms of phytoplankton as survival afterlives in an unstable environment. Oceanologica Acta 1:493-509.

Marquet PA, Abades SR, Labra FA (2007) Biodiversity power laws. Scaling Biodiversity, eds Marquet PA, Brown JH, pp 441-461.

Marquet PA, Quiñones RA, Abades S, Labra F, Tognelli M, Arim M, Rivadeneira M (2005) Scaling and power-laws in ecological systems. J Exp Biol 208:1749-1769.

Martiny JBH, Bohannan BJM, Brown JH, et al. (2006) Microbial biogeography: putting microorganisms on the map. Nature Rev Microbiol 4:102-112.

Martiny JBH, Eisen JA, Penn K, Allison SD, Horner-Devine MC (2011) Drivers of bacterial β-diversity depend on spatial scale. Proc Natl Acad Sci USA 108:7850-7854.

May R (1975) Ecology and Evolution of Communities (Harvard Univ. Press, Cambridge, MA), pp 81-120.

Mayhew PJ, Bell MA, Benton TG, McGowan AJ (2012) Biodiversity tracks temperature over time. Proc Natl Acad Sci USA 9(38):15141-15145.

McGill BJ, Etienne RS, Gray JS, Alonso A, Anderson MJ, Benecha HK, Enquist BJ, Green JL, He F, Hurlbert AH, Magurran AE, Marquet PA, Maurer BA, Ostling A, Soykan CU, Ugland KI, White EP (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. Ecol Lett 10:995-1015.

Medlin LK, Elwood HJ, Stickel S, Sogin ML (1991) Morphological and genetic variation within the diatom *Skeletonema costatum* (*Bacillariophyceae*): Evidence for a new species, *Skeletonema pseudocostatum*. J Phycol 27:514-524.

Medlin LK, Jung I, Bahulikar R, Mendgen K, Kroth P, Kooistra WHCF (2008). Evolution of the Diatoms. VI. Assessment of the new genera in the araphids using molecular data. Nova Hedwigia Beih

133:81-100.

Meusnier I, Singer GAC, Landry JF, Hickey DA, Hebert PDN, Hajibabaei M (2008) A universal DNA mini-barcode for biodiversity analysis. BMC Genomics 9(1):214.

Miller CB (2004) Biological Oceanography (Oxford: Blackwell Publishing), 402 pp.

Mitzenmacher M (2004) A brief history of generative models for power law and lognormal distributions. Internet Math 1:226-251.

Mitzenmacher M (2006) The future of power law research. Internet Math 2:525-534.

Moore LA (2010) Niche differentiation, rarity, and commonness in the sympatric Australian white-tailed rats: Uromys caudimaculatus and Uromys hadrourus. PhD thesis, James Cook University.

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How many species are there on Earth and in the ocean?, PLoS Biol 9:e1001127 .

Moriarty R, Buitenhuis ET, Le Quéré C, Gosselin M-P (2013) Distribution of known macrozooplankton abundance and biomass in the global ocean. Earth Syst Sci Data 5:241-257.

Morlon H, Chuyong G, Condit R, Hubbell S, Kenfack D, Thomas D, Valencia R, Green JL (2008) A general framework for the distance-decay of similarity in ecological communities. Ecol Lett 11:904-917.

Morton RD, Law R (1997) Regional species pools and the assembly of local ecological communities. J Theor Biol 187:321-331.

Mouquet N, Loreau M (2002) Coexistence in metacommunities: The regional similarity hypothesis. Am Nat 159:420-426.

Naeem S, Thompson LJ, Lawler SP, Lawton JH, Woodfin RM (1994) Declining biodiversity can alter the performance of ecosystems. Nature 368:734-737.

Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) (2013) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 5-24.

Nanjappa D, Audic S, Romac S, Kooistra WH, Zingone A (2014) Assessment of species diversity and distribution of an ancient diatom lineage using a DNA metabarcoding approach. PloS one 9(8):e103810.

Negro AI, De Hoyos C (2005) Relationships between diatoms and the environment in Spanish reservoirs. Limnetica 24(1-2):133-144.

Nekola JC, Šizling AL, Boyer AG, Storch D (2008) Artifactions in the Log-Transformation of Species Abundance Distributions. Folia Geobot 43:259-268.

Nekola JC, White PS (1999) The distance decay of similarity in biogeography and ecology. J Biogeogr 26:867-878.

Nelson DM, Treguer P, Brzezinski MA, Leynaert A, Queguiner B (1995) Production and dissolution of biogenic silica in the ocean - revised global estimates, comparison with regional data and relationship to biogenic sedimentation. Global Biogeochem Cy 9:359-372.

Newman MEJ (2000) The power of design. Nature 405:412.

Newman MEJ (2005) Power laws, Pareto distributions and Zipf's law. Contemp Phys 46(5):323-351.

Nolte V, Pandey RV, Jost S, Medinger R, Ottenwälder B, Boenigk J, et al. (2010) Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. Mol Ecol 19:2908-2915.

Oksanen J (2011) Multivariate analysis of ecological communities in R: vegan tutorial.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, et al. (2013) vegan: Community Ecology Package. R package version 2.0-10.

Oliver MJ, Irwin AJ (2008) Objective global ocean biogeographic provinces. Geophys Res Lett 35:L15601.

Olszewski TD (2004) A unified mathematical framework for the measurement of richness and evenness within and among multiple communities. Oikos 104:377–387.

Patrick R (1971) The effects of increasing light and temperature on thestructure of diatom communities. Limnol Oceanogr 16:405–421.

Palmer MW, McGlinn DJ, Fridley JD (2008) Artifacts and artifictions in biodiversity research. Folia Geobot 43(3):245-257.

Pan YD, Stevenson RJ, Hill BH, Herlihy AT (2000) Ecoregions and benthic diatom assemblages in Mid-Atlantic Highlands streams, USA. J N Am Benthol Soc 19:518–40.

Pawlowski J, et al. (2012) CBOL protist working group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. PLoS Biol 10:e1001419.

Pedrós-Alió C (2006) Marine microbial diversity: can it be determined? Trends Microbiol 14:257-263.

Pedrós-Alió C (2012) The Rare Bacterial Biosphere. Annu Rev Marine Sci 4:449-466.

Pereira PRG, Fuchs BM, Alonso C, Oliver MJ, van Beusekom JEE, Amann R (2010) Distinct flavobacterial communities in contrasting water masses of the North Atlantic Ocean ISMEJ 4: 472-487.

Peres-Neto P, Legendre P, Dray S, Borcard D (2006) Variation partitioning of species data matrices: estimation and comparison of fractions. Ecology 87:2614-2625.

Pesant et al., (2015) Open science resources for the discovery and analysis of the Tara Oceans Data Collection. Sci Data (In press)

Peters RH (1983) The Ecological Implications of Body Size (Cambridge Univ. Press, Cambridge).

Peterson RJ, Stramma L (1991) Upper level circulation in the South Atlantic Ocean. Prog Oceanogr 26(1):1-73.

Pickett-Heaps JD, Pickett-Heaps J (2003) Diatoms: Life in Glass Houses. Cytographics.

Plotkin JB , Potts MD, Leslie N, Manokaran N, LaFrankie J, Ashton PS (2000) Species-area curves,

spatial aggregation, and habitat specialization in tropical forests. J Theor Biol 207:81-99.

Pomeau Y, Manneville P (1980) Intermittent transition to turbulence in dissipative dynamical systems. Commun. Math Phys 14:189-197.

Post WM, Pimm SL (1983) Community assembly and food web stability. Math Biosci 64:169-192.

Potapova MG, Charles DF (2003) Distribution of benthic diatoms in U.S. rivers in relation to conductivity and ionic composition. Freshwater Biol 48:1311–28.

Preston FW (1948) The commonness and rarity of species. Ecology 29:254-283.

Pueyo S (2006) Diversity: between neutrality and structure. Oikos 112:392-405.

Qian H (2009) Beta diversity in relation to dispersal ability for vascular plants in North America. Global Ecol Biogeogr 18:327-332.

Quijano-Scheggia SI, Garcés E, Lundholm N, Moestrup O, 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. Phycologia 48(6):492-509.

R Development Core Team (2009) R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Rabosky DL, Sorhannus U (2009) Diversity dynamics of marine planktonic diatoms across the Cenozoic. Nature 457:183-187.

Raes J, Letunic I, Yamada T, Jensen LJ, Bork P (2011) Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. Mol Syst Biol 7:473.

Raghukumar S (2002) Ecology of the marine protists, the *Labyrinthulomycetes* (*Thraustochytrids and Labyrinthulids*). Eur J Protistol 38:127-145.

Ramette A, Tiedje JM (2007) Biogeography: an emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. Microb Ecol 53:197-207.

Rampen SW, Schouten S, Elda Panoto F, Brink M, Andersen RA, Muyzer G, Abbas B, Sinninghe Damsté JS (2009) Phylogenetic position of *Attheya longicornis* and *Attheya septentrionalis* (bacillariophyta). J Phycol 45:444-453.

Reuter JS, Mathews DH (2010) RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics 11:129.

Reyes-Prieto A, Hackett JD, Soares MB, Bonaldo MF, Bhattacharya D (2006) Cyanobacterial contribution to algal nuclear genomes is primarily limited to plastid functions. Curr Biol 16:2320-2325.

Riaz T, Shehzad W, Viari A, Pompanon F, Taberlet P, Coissac E (2011) ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. Nucleic Acids Res 39(21):e145.

Ribera d'Alcala M, Conversano F, Corato F, Licandro P, Mangoni O, Marino D, Mazzocchi MG, Modigh M, Montresor M, Nardella M, Saggiomo V, Sarno D, Zingone A (2004) Seasonal patterns in

plankton communities in a pluriannual time series at a coastal Mediterranean site (Gulf of Naples): an attempt to discern recurrences and trends. Sci Mar 68 (Suppl. 1):65-83.

Ricklefs RE, Schluter D (1993) Species Diversity in Ecological Communities (Univ. Chicago Press, Chicago).

Rodríguez-Martínez R, Gabrielle R, Guillem S, Ramon M (2013) Biogeography of the uncultured marine picoeukaryote MAST-4: temperature-driven distribution patterns. ISMEJ 7:531-1543.

Rombouts I, Beaugrand G, Ibanez F, Gasparini S, Chiba S, Legendre L (2009) Global latitudinal variations in marine copepod diversity and environmental factors. Proc R Soc B 276:3053-3062.

Roque FO, Siqueira T, Bini LM, Ribeiro MC, Tambosi LR, Ciocheti G, Trivinho-Strixino S (2010) Untangling associations between chironomid taxa in Neotropical streams using local and landscape filters. Freshwater Biol 55(4):847-865.

Round FE, Crawford RM, Mann DG (1990) The diatoms: biology and morphology of the Genera (Cambridge University Press, Cambridge), pp747.

Roxburgh SH, Chesson P (1998) A new method for detecting species associations with spatially autocorrelated data. Ecology 79(6):2180–2192.

Roy K, Jablonski D, Valentine JW, Rosenberg G (1998) Marine latitudinal diversity gradients: tests of causal hypotheses. Proc Natl Acad Sci USA 95:3699- 3702.

Rummel JD, Roughgarden J (1985) A theory of faunal buildup for competition communities. Evolution 39:1009-1033.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S et al. (2007) The Sorcerer II global ocean sampling expedition: northwest atlantic through eastern tropical pacific. PLoS Biol 5:e77.

Rutherford S, D'Hondt S, Prell (1999) W Environmental controls on the geographic distribution of zooplankton diversity. Nature 400:749 -753.

Rynearson TA, Armbrust EV (2000) DNA fingerprinting reveals extensive genetic diversity in a field population of the centric diatom *Ditylum brightwellii*. Limnol Oceanogr 45:1329-1340.

Sabater S, Roca JR (1992) Ecological and biogeographical aspects of diatom distribution in Pyrenean springs. Br Phycol J 27:203-213.

Scanlan DJ, Ostrowski M, Mazard S, Dufresne A, Garczarek L, Hess WR, Post AF, Hagemann M, Paulsen I, Partensky F (2009) Ecological genomics of marine *Picocyanobacteria*. Microbiol Mol Biol Revi 73:249-299.

Scherer-Lorenzen M (2005) Biodiversity and ecosystem functioning: basic principles. Biodiversity: Structure and Function. Encyclopedia of Life Support Systems (EOLSS), eds Barthlott W, Linsenmair KE, Porembski S, (EOLSS Publisher, Oxford).

Schnepf E (1994) Light and electron microscopical observations in *Rhynchopus coscinodiscivorus* sp. nov., a Colorless, phagotrophic *Euglenozoon* with concealed flagella. Arch für Protistenkd. 144:63-74.

Shannon CE, Weaver W (1949) The mathematical theory of communication. Urbana, University of

Illinois Press.

Shurin JB, Cottenie K, Hillebrand H (2009) Spatial autocorrelation and dispersal limitation in freshwater organisms. Oecologia 159:151-159.

Siedler G, Griffies S, Gould J and Church J (2013) Ocean Circulation and Climate: A 21st century perspective (Academic Press), 904 pp.

Signorini SR, McClain CR (2009) Environmental factors controlling the Barents sea spring-summer phytoplankton blooms. Geophys Res Lett 36:L10604.

Sims PA, Mann DG, Medlin (2006) LK Evolution of the diatoms: insights from fossil, biological and molecular data. Phycologia 45:361-402.

Sims PA, Mann DG, Medlin LK (2006) Evolution of the diatoms: insights from fossil, biological and molecular data. Phycologia 45(4):361-402.

Siokou-Frangou I, Christaki U, Mazzocchi MG, Montresor M, Ribera d'Alcalá M, Vaqué D, Zingone A (2010) Plankton in the open Mediterranean Sea: a review. Biogeosciences 7(5):1543-1586.

Skov J, Lundholm N, Pocklington R, Rosendahl S, Moestrup O (1997) Studies on the marine planktonic diatom *Pseudo-nitzschia*. 1. Isozyme variation among isolates of *P. pseudodelicatissima* during a bloom in Danish coastal waters. Phycologia 36:374-380.

Skovgaard A, Saiz E (2006) Seasonal occurrence and role of protistan parasites in coastal marine zooplankton. Mar Ecol Prog Ser 327:37-49.

Slamovits CH, Okamoto N, Burri L, James ER, Keeling PJ (2011) A bacterial proteorhodopsin proton pump in marine eukaryotes. Nat Commun 2:1-6.

Smalley GW, Coats DW (2002) Ecology of the red-tide dinoflagellate ceratium furca: distribution, mixotrophy, and grazing Impact on ciliate populations of chesapeake bay. J Eukaryot Microbiol 49:63-73.

Smetacek V (1998) Diatoms and the silicate factor. Nature 391:224-225.

Smetacek V, Klaas C, Strass VH, Assmy P, et al. (2012) Deep carbon export from a Southern Ocean iron-fertilized diatom bloom. Nature 487:313-319.

Smith AB, McGowan AJ (2005) Cyclicity in the fossil record mirrors rock outcrop area. Biol lett 1(4):443-445.

Smith MA, Fisher BL, Hebert PDN (2005) DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. Phil Trans R Soc B 360(1462):1825-1834.

Smol JP, Stoermer EF (2010) The Diatoms: Applications for Environmental and Earth Sciences, Second Edition (Cambridge University Press), pp 667.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ (2006) Microbial diversity in the deep sea and the underexplored rare biosphere. Proc Natl Acad Sci USA 103(32):12115-12120.

Soininen J (2002) Responses of epilithic diatom communities to environmental gradients in some Finnish rivers. Internat Rev Hydrobiol 87:11–24.

Soininen J, Kokocinski M, Estlander S, Kotanen J, Heino J (2007) Neutrality, niches, and determinants of plankton metacommunity structure across boreal wetland ponds. Ecoscience 14(2):146-154.

Soininen J, McDonald R, Hillebrand H (2007) The distance decay of similarity in ecological communities. Ecography 30:3-12.

Soininen J, Paavola R, Muotka T (2004) Benthic diatom communities in boreal streams: community structure in relation to environmental and spatial gradients. Ecography 27:330-342.

Sorhannus U (2007) A nuclear-encoded small-subunit ribosomal RNA timescale for diatom evolution. Mar Micropaleontol 65:1-12.

Sornette D (2006) Critical Phenomena in Natural Sciences, Second edition (2nd printing), pp 527.

Soudek DJR, Robinson GGC (1983) Electrophoreticanalysis of the species and population structure of the diatom *Asterionella formosa*. Can J Bot 61:418-433.

Sournia A, Chrdtiennot-Dinet MJ, Ricard M (1991) Marine phytoplankton: how many species in the world ocean? J Plankton Res 13 (5):1093-1099.

Stachowicz JJ, Best RJ, Bracken MES, Graham M (2008) Complementarity in marine biodiversity manipulations: reconciling divergent evidence from field and mesocosm experiments. Proc Natl Acad Sci USA 105:18842-18847.

Steele PR, Pires JC (2011) Biodiversity assessment: state-of-the-art techniques in phylogenomics and species identification. Am J Bot 98(3):415-425.

Stemmann L, Youngbluth M, Robert K, Hosia A, Picheral M, Paterson H, Ibanez F, Guidi L, Lombard F, Gorsky G (2008) Global zoogeography of fragile macrozooplankton in the upper 100-1000 m inferred from the underwater video profiler. ICES J Mar Sci 65: 433-442.

Stevens GC (1989) The latitudinal gradient in geographical range: how so many species co-exist in the tropics. Am Nat 133:240-256.

Stevenson RJ (1997) Scale-dependent determinants and consequences of benthic algal heterogeneity. J N Am Benthol Soc 16:248-262.

Stoermer EF, Smol JP (1999) The Diatoms: Applications for the Environmental and Earth Sciences. (Cambridge University Press, Cambridge), pp 469.

Storch D, Gaston KJ (2004) Untangling ecological complexity on different scales of space and time. Basic Appl Ecol 5: 389-400.

Stumpf MPH, Porter MA (2012) Critical truths about power laws. Science 335(6069):665-666.

Sul WJ, Oliver TA, Ducklow HW, Amaral-Zettler LA, Sogin ML (2013) Marine bacteria exhibit a bipolar distribution. Proc Natl Acad Sci USA 110(6):2342-7.

Sunagawa S, et al. (2015) Structure and Function of the Global Ocean Microbiome. Science (in press).

Suttle CA (2007) Marine viruses - major players in the global ecosystem. Nat Rev Micro 5:801-812.

Swanberg NR (1974) Thesis. Massachussetts Institute of Technology & Woods Hole Oceanographic Institution, Woods Hole, USA.

Taberlet P, Coissac e, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. Mol Ecol 21:2045-2050.

Taberlet P, Prud'homme SM, Campione E, Roy J, Miquel C, Shehzad W, Gielly L, Rioux D, Choler P, Clément JC, Melodelima C, Pompanon F, Coissac E (2012) Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. Mol Ecol 21:1816-1820.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol Biol Evol 30:2725-2729.

Taylor FJR, Pahlinger U (1987) Ecology of dinoflagellates. The biology of dinoflagellates, ed Taylor FJR. Botanical Monographs 21:399-529.

Taylor HR, **Harris WE** (2012) An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. Mol Ecol Res 12:377-388.

Temperton B, Gilbert JA, Quinn JP, McGrath JW (2011) Novel analysis of oceanic surface water metagenomes suggests importance of polyphosphate metabolism in oligotrophic environments. PLoS one 6:e16499.

ter Braak CJF (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67(5):1167-1179.

ter Braak CJF, Prentice IC (1988) A theory of Gradient Analysis. Adv Ecol Res 18:271-317.

Thurman HV (1997) Introductory Oceanography (New Jersey, USA).

Tilman D (2004) A stochastic theory of resource competition, community assembly and invasions. Proc Natl Acad Sci USA 101:10854-10861.

Tittensor DP, Mora C, Jetz W, Lotze HK, Ricard D, et al. (2010) Global patterns and predictors of marine biodiversity across taxa. Nature 466:1098-1101.

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 33:2-22.

Tuomisto H (2010b) A diversity of beta diversities: straightening up a concept gone awry. Part 2. Quantifying beta diversity and related phenomena. Ecography 33:23-45.

Turner JRG, Gatehouse CM, Corey CA (1987) Does solar energy control organic diversity? Butterflies, moths and the British climate. Oikos 48:195-205.

Turner JT (2004) The importance of small planktonic copepods and their roles in pelagic marine food webs. Zool Stud 43:255-226.

Valentini A, Pompanon F, Taberlet P (2009) DNA barcoding for ecologists. Trends Ecol Evol 24(2):110-117.

Van der Gucht K, Cottenie K, Muylaert K, Vloemans N, Cousin S, et al. (2007) The power of species sorting: Local factors drive bacterial community composition over a wide range of spatial scales. Proc Natl Acad Sc USA 104:20404-20409.

Vanormelingen P, Evans KM, Chepurnov VA, Vyverman W, Mann DG (2013) Molecular species discovery in the diatom *Sellaphora* and its congruence with mating trials. Fottea Olomouc 13(2):133-148.

Vanormelingen P, Verleyen E, Vyverman W (2008) The diversity and distribution of diatoms: from cosmopolitanism to narrow endemism. Biodiver Conserv 17:393-405.

Vellend M (2010) Conceptual synthesis in community ecology. Q Rev Biol 85:183-206.

Venables WN, Ripley BD (2002) Modern Applied Statistics with S, Fourth edition (Springer).

Villar E, Farrant GK, Follows M, Garczarek L et al. (2015) Environmental disturbance in Agulhas rings affect inter-ocean plankton dispersal. Science (In press).

Volkov I, Banavar JR, Hubbell SP, Maritan A (2003) Neutral theory and relative species abundance in ecology. Nature 424:1035-1037.

Weinbauer MG, Rowe JM, Wilhelm SW (2010) Determining rates of virus production in aquatic systems by the virus reduction approach. Manual of Aquatic Viral Ecology, eds Wilhelm SW, Weinbauer MG and Suttle CA (ASLO) p. 1-8.

Werner D (1977) The Biology of Diatoms, Botanical Monographs, University of California Press, volume 13:498.

Westoby M (2006) Phylogenetic ecology at world scale, a new fusion between ecology and evolution. Ecology 87:163-165.

Wetzel CE, Bicudo DdC, Ector L, Lobo EA, Soininen J, et al. (2012) Distance decay of similarity in neotropical diatom communities. PLoS one 7(9):e45071.

Whitaker R J, Grogan DW, Taylor JW (2003) Geographic barriers isolate endemic populations of hyperthermophilic archaea. Science 301:976-978.

Whittaker R (1962) Classification of natural communities. Bot Rev 28:1-239.

Whittaker RH (1960) Vegetation of the Siskiyou Mountains, Oregon and California. Ecol Monogr 30:279-338.

Whittaker RH (1965) Dominance and diversity in land plant communities. Science 147:250-260.

Whittaker RH (1972) Evolution and measurement of species diversity. Taxon 21:213-251.

Whittaker RH (1975) Communities and ecosystems, Second edition (Macmillan, New York).

Wiebe PH, Bucklin A, Madin L, Angel MV, Sutton T, Pagés F, Hopcroft RR, Lindsay D (2010) Deep-sea sampling on CMarZ cruises in the Atlantic Ocean - an Introduction. Deep-Sea Res II 57:2157-2166.

Will KW, Mishler BD, Wheeler QD (2005) The perils of DNA barcoding and the need for integrative taxonomy. Syst Biol 54(5):844-851.

Worden AZ, Not F (2008) Ecology and diversity of picoeukaryotes. Microbial Ecology of the Ocean, ed Kirchman DL (Wiley), pp 159-205.

Wu J, Jones KB, Li H, Loucks OL (Eds.) (2006) Scaling and Uncertainty Analysis in Ecology: Methods and Applications XVIII, pp 338.

Yao M, Li Y-L, Yang X-D, Liu Q (2011) Three-year changes in planktonic diatom communities in a eutrophic lake in Nanjing, Jiangsu Province, China. J Freshwater Ecol 26(1):133-141.

Yoccoz NG (2012) The future of environmental DNA in ecology. Mol Ecol 21:2031-2038.

Yool A, Tyrrell T (2003) Role of diatoms in regulating the ocean's silicon cycle. Global Biogeochemical Cycles 17 (4): 1103–1124.

Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattachrya DA (2004) molecular timeline for the origin of photosynthetic eukaryotes. Mol Biol Evol 21:809-818.

Yu DW, Ji Y, Emerson BC, Wang X, Ye C, Yang C, Ding Z (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods Ecol Evol 3:613-623.

Yusupov MM, Yusupova GZ, Baucom A, et al. (2001) Crystal structure of the ribosome at 5.5 A resolution. Science 292 (5518):883-96.

Zagoskin MV, Lazareva VI, Grishanin AK, Mukha DV (2014) Phylogenetic Information Content of Copepoda Ribosomal DNA Repeat Units: ITS1 and ITS2 Impact. Biomed Res Int 18:926342.

Zelinka M, Marvan P (1961) Zur Pr¨azisierung der biologischen Klassifikation des Reinheit fliessender Gew¨asser. Archiv f¨urHydrobiologie 57:389–407.

Zingone A, Sarno D (2001) Recurrent patterns in coastal phytoplankton from the Gulf of Naples. Arch Oceanograf Limnol 22:13-118.

ANNEXES

Summary

A. Glossary

Alpha diversity: The diversity within a particular area or ecosystem; usually expressed by the number of species (i.e., species richness) in that ecosystem.

Beta diversity: A comparison of diversity between ecosystems, usually measured as the amount of species change between the ecosystems.

Gamma diversity: A measure of the overall diversity within a large region.

Complex systems: systems with a large number of mutually interacting parts, often open to their environment, which self-organize their internal structure and their dynamics with novel and sometimes surprising macroscopic "emergent" properties.

Criticality (in physics): a state in which spontaneous fluctuations of the order parameter occur at all scales, leading to diverging correlation length and susceptibility of the system to external influences

Power law distribution: a specific family of statistical distribution appearing as a straight line in a log-log plot; does not possess characteristic scales and exhibit the property of scale invariance.

Hellinger transformation: Relativization by row (sample unit) totals, followed by taking the square root of each element in the matrix.

Distance Decay: the property by which two nearby points have more similar characteristics than two distant points.

Distance Matrix: A square and (usually) symmetric matrix in which the rows and the columns represent (usually) samples. The entries represent some index of the difference between samples; the measure could be Euclidean distance, Manhattan (City Block) Distance, Bray-Curtis dissimilarity, the Jaccard coefficient, or any of a huge number of possibilities. The diagonal elements (the difference between a sample and itself) is usually zero.

Environmental Gradient: a spatially varying aspect of the environment which is expected to be related to species composition.

Jackknife: A (usually) computer-intensive method to estimate parameters, and/or to gauge uncertainty in these estimates. The name is derived from the method that each observation is removed (i.e. cut with the knife) one at a time (or two at a time for the second-order Jackknife, and so on) in order to get a feeling for the spread of data.

Stress: A measure of the optimality of an ordination solution (i.e. the relationship between the similarity in species composition and the closeness in ordination space), used as part of the algorithm of NMDS.

Explained variance: Share of the total variance which is accounted for by the model. Explained variance is computed as the complement to residual variance, divided by total variance. It is expressed as a percentage.

Multivariate analysis (MVA): It is based on the statistical principle of multivariate statistics, which involves observation and analysis of more than one statistical variable at a time. In design and analysis, the technique is used to perform trade studies across multiple dimensions while taking into account the effects of all variables on the responses of interest.

Principal component (PC): Principal Components (PCs) are composite variables, i.e. linear functions of the original variables, estimated to contain, in decreasing order, the main structured information in the data. A PC is the same as a score vector, and is also called a latent variable or a factor.

P-value: The p-value measures the probability that a parameter estimated from experimental data should be as large as it is, if the real (theoretical, non-observable) value of that parameter were actually zero. Thus, p-value is used to assess the significance of observed effects or variations: a small p-value means a small risk of mistakenly concluding that the observed effect is real. The usual limit used in the interpretation of a p-value is 0.05 (or 5%). If p-value < 0.05, the observed effect can be presumed to be significant and is not due to random variations.

Regression coefficient: In a regression model equation, regression coefficients are the numerical coefficients that express the link between variation in the predictors and variation in the response.

Residual: A measure of the variation that is not taken into account by the model. The residual for a given sample and a given variable is computed as the difference between observed value and fitted (or projected, or predicted) value of the variable on the sample.

R-square: The R-square of a regression model is a measure of the quality of the model. Also known as coefficient of determination, it is computed as 1 - (Residual Y-variance), or (Explained Y-variance)/100. For Calibration results, this is also the square of the correlation coefficient between predicted and measured values, and the R-square value is always between 0 and 1. The closer to 1, the better. The R-square is displayed among the plot statistics of a Predicted vs. Reference plot. When based on the calibration samples, it tells about the quality of the fit. When computed from the validation samples (similar to the "adjusted R^2" found in the literature) it tells about the predictive ability of the model.

r selection: Selection of life-history traits which promote an ability to multiply rapidly in numbers - the traits being, broadly, small size, precocious reproduction, a large reproductive allocation and the production of many but small offspring.

Realized niche: That portion of its fundamental niche occupied by a species when competitors or predators are present

Rarefaction curve: The statistical expectation of the number of species in a survey or collection as a function of the accumulated number of individuals or samples, based on resampling from an observed sample set.

B. Diversity and similarity Indices

Species richness. Species richness is the number of different species present in an area. The more species present in a sample the 'richer' the area. Species richness as a measure on its own takes no account of the number of individuals of each species present. It gives equal weight to those species with very few individuals and those with many individuals. A better measure of diversity should take into account the abundance of each species.

Shannon-Wiener index. The Shannon diversity index (H) is another index that is commonly used to characterize species diversity in a community. Like Simpson's index, Shannon's index accounts for both abundance and evenness of the species present. The formula for calculating H is presented as:

$$H = -\sum_{i=1}^{s} p_i \ln p_i$$

where, H = the Shannon diversity index

p_i = the proportion of species i relative to total number of species present.

s = numbers of species encountered

The fact that the index incorporates both components of biodiversity can be seen as both a strength and a weakness. It is a strength because it provides a simple, synthetic summary, but it is a weakness because it makes it difficult to compare communities that differ greatly in richness

Evenness/ Shannon's equitability. Shannon's equitability (E_H) can be calculated as,

$$E_H = H / H_{max}$$

Equitability assumes a value between 0 and 1 with 1 being complete evenness.

The effective number of species (H_1). The effective number of species or true diversity, or, refers to the number of equally-abundant types needed for the average proportional abundance of the types to equal that observed in the dataset of interest (where all types may not be equally abundant). Shannon index can be converted to true diversities using the formulae,

$$\mathbf{H_1} = \exp(SDI),$$

Where, SDI is the value of Shannon index.

Jaccard's index. A dissimilarity measure that applies to samples of presence/absence data, similar to the matching coefficient, but ignoring the number of co-absences between two samples. Jaccard's index is the simplest summary of this, taking the following form:

$$J = \frac{S_a}{S_a + S_b + S_c}$$

Where, S_a and S_b are the numbers of species unique to samples a and b, respectively, S_c is the number of species common to the two samples. It only utilizes the richness component of diversity and thus is simply the fraction of species shared between the samples.

Bray Curtis dissimilarity. A measure of dissimilarity commonly used in ecology, to measure differences between multivariate samples of species abundances or biomasses. The general formula for calculating the Bray-Curtis dissimilarity between samples '*a'* and '*b'* is as follows:

$$\mathrm{B_{a,b}} = \frac{\sum_{j=1}^{J}(n_{aj} - n_{bj})}{\mathrm{n_{a+} + n_{b+}}}$$

Where n_{aj} and n_{bj} are the numbers of species unique to samples a and b, respectively

n_{a+} and n_{b+} are the sample (row) totals.

One of the assumptions of the Bray-Curtis measure is that the samples are taken from the same physical size, be it area or volume. This is because dissimilarity will be computed on raw counts, not on relative counts. This measure takes on values between 0 (samples identical) and 1 (samples completely disjoint). If the Bray-Curtis dissimilarity is subtracted from 1, a measure of similarity is obtained, called the Bray-Curtis index.

Euclidean distances. A distance measure between vectors where squared differences between corresponding elements are summed, followed by taking the square root of this sum. The general formula for calculating the Euclidean distance between samples '*a'* and '*b'* is as follows:

$$\mathrm{d_{(a,b)}} = \sqrt{\sum_{i=1}^{J}(a_i - b_i)^2}$$

C. Multivariate statistical methods

Non-metric multidimensional scaling (NMDS). Maximizes rank-order correlation between distance measures and distance in ordination space. Points are moved to minimize "stress". Stress is a measure of the mismatch between the two kinds of distance.

Principal components analysis (PCA). PCA is a projection of sample points from the multidimensional variable space onto a 'best-fitting' plane or other low-dimensional solution. Linear representation of the data: usually inadequate for community analysis but good for reducing environmental measurements and to detect patterns in the distribution of environmental variables. PC axes reflect the sum of contributions from each of the environmental parameters and represent simple linear combinations of the abiotic variables, their coefficients termed eigenvectors. Dissimilarity between two samples j and k is defined as their Euclidean distance apart in the multidimensional space.

Redundancy analysis (RDA). Extension of principal component analysis to include external explanatory variables; the solution is constrained to have dimensions that are linearly related to these explanatory variables.

Mantel test. It computes a correlation between two *n* by *n* distance matrices. The null hypothesis is that the observed relationship between the two distance matrices could have been obtained by any random arrangement in space (or time, or treatment assignment) of the observations through the study area. The null hypothesis is no relationship between the two distance matrices. It calculates a rank correlation coefficient between all the elements of their respective similarity or dissimilarity matrices. Thus, if the among-sample relationships are the same, then the rank correlation R = 1, a perfect match.

Partial Mantel Test. It allows a comparison to be made among two variables while controlling for the third.

Multiple regression on distance matrices (MRM). MRM involves a multiple regression of a response matrix on any number of explanatory matrices, where each matrix contains distances or similarities (in terms of ecological, spatial, or other attributes) between all pair-wise combinations of n objects (sample units); tests of statistical significance are performed by permutation.

Analysis of similarities (ANOSIM). An analysis that compares several groups, usually their mean values, can also be thought of as a variant of regression analysis when the independent variable is a categorical variable. If $\bar{r}W$ is defined as the average of all rank similarities among replicates within samples, and $\bar{r}B$ is the average of rank similarities arising from all pairs of replicates between different groups of samples, then

$$R = \frac{(\bar{r}_B - \bar{r}_W)}{\frac{1}{2}M}$$

where M = n(n-1)/2 and n is the total number of samples under consideration. R = 1 only if all replicates within sites are more similar to each other than any replicates from different sites. If R is approximately zero, similarities between and within sites will be the same on average.

Cluster analysis. A group-forming technique, which constructs groups of samples (or variables) that have high internal similarity, while maintaining low similarity between groups.

Ward clustering. A specific hierarchical clustering algorithm which minimizes the within-cluster inertia at each clustering step, equivalent to maximizing the between-cluster inertia.
Variation Partitioning.

Linear and multiple regression. An approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variable) denoted X. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.

Maximum likelihood estimation. A method of estimating population characteristics from a sample by choosing the values of the parameters that will maximize the probability of getting the particular sample actually obtained from the population.

Wilcoxon signed-ranks method. It tests the null hypothesis that two related medians are the same. This procedure allows testing for differences between paired scores of two related samples when the assumptions required by the paired-samples t test are not met. Ranks are based on the absolute value of the difference between the two test variables.

Mann-Whitney U Test. This tests the differences between the two approaches by analyzing differences in the absolute values of the abundances measured. The test is based on the null hypothesis that two independent samples come from the same population and does not assume normality or equal variances in the data.

D. Resolution of V9 18S rDNA tags in diatom phylogeny

Diatoms are traditionally divided into two orders: the Centrales and the Pennales. Of these, the former is considered paraphyletic to the later. Further, recent classification efforts, based on ultra-stuctural details of the frustule, have divided the diatoms into four “groups of convenience”: (i) polar centric (polar Coscinodiscophyceae), (ii) radial centric (polar Coscinodiscophyceae), (iii) araphid pennate (Fragilariophyceae) and (iv) raphid pennate (Bacillariophyceae). **Figure D1** illustrates the accepted relationships between diatom classes based on Bayesian inference (Kooistra et al., 2003; 2007; Medlin et al., 2008; Rampen et al., 2009). The phylogeny was obtained using 18S rDNA sequences longer than 1,500 nucleotides. These were aligned to sequences stored in the ARB database. The consensus trees was constructed using the program *MrBayes* 3.1.2. (Huelsenbeck et al. 2001), using the general-time-reversible (GTR) model with gamma-distributed rate variation across sites and a proportion of invariable sites (for detail, see Kooistra et al. 2003).

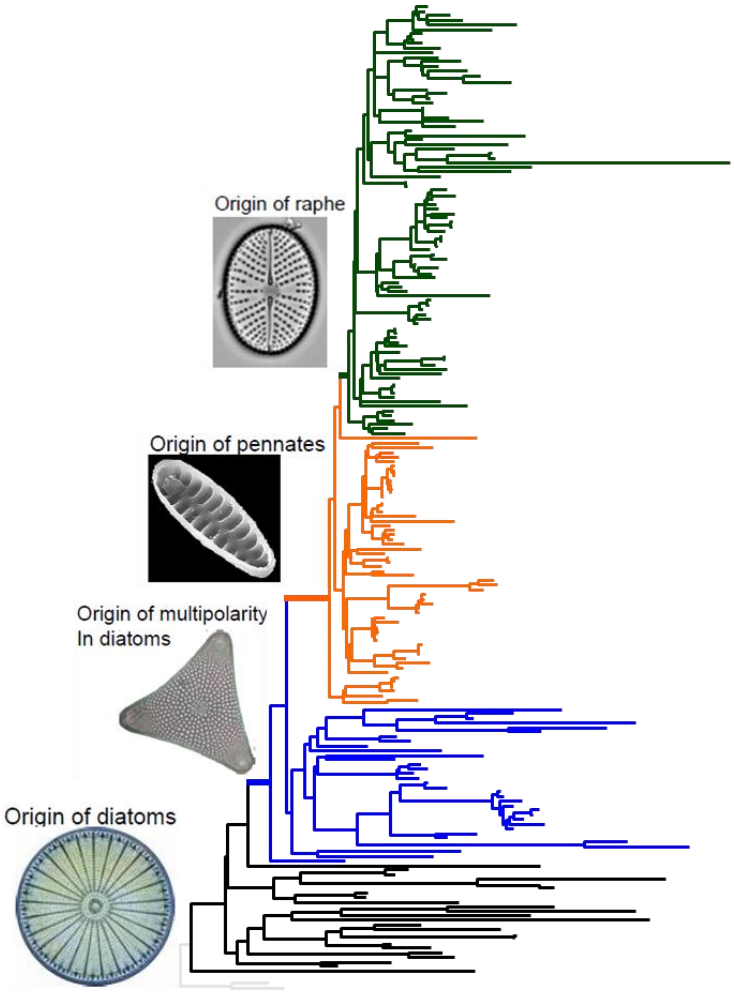


Figure D1. Diatom phylogeny inferred with Bayesian inference analyses of 18S rDNA (modified from Kooistra et al., 2003). A principal dichotomy was revealed showing a clade of radial centrics (basal clade) and another with multipolar centrics and pennates. Centric diatoms are characterized by a typically circular to elliptical or polygonal valves. Araphid pennates are characterized by elongated valves while raphid pennates with the raphe slit. Green = raphid pennates, orange = araphid pennates, blue = bi (multi) polar centrics, black = radial centrics and grey = outgroup species.

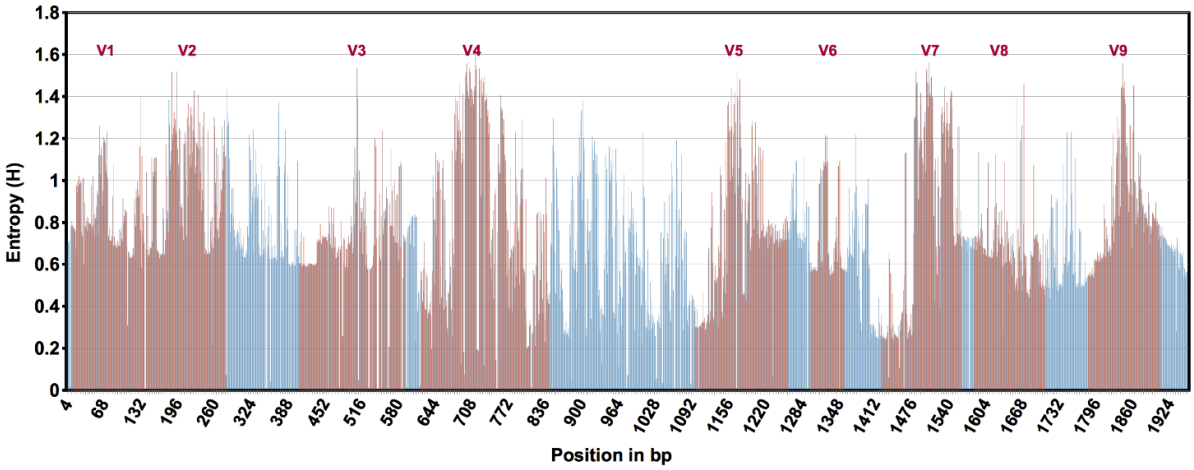


Figure D2. Bar plot illustrating the Shannon entropy associated with each position along an alignment of 2497 diatom 18S rDNA sequences. Bars colour-coded in red correspond to positions in the V1-V9 regions of the gene. Blue color is conserved region.

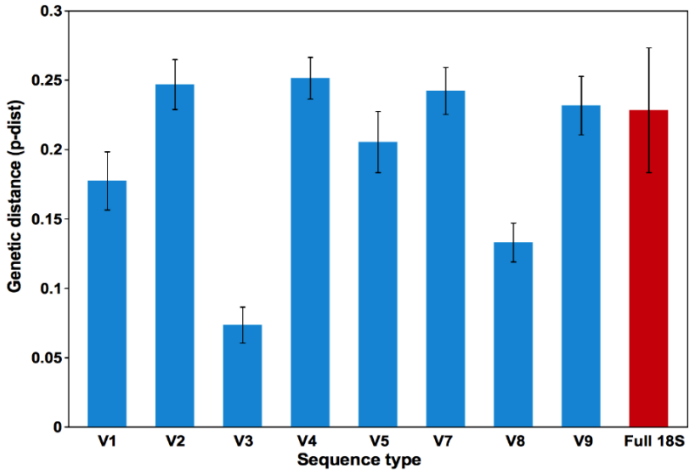


Figure D3. 18S rDNA based divergence determination of Bacillariophyta.

Table 1.1. Hypervariable region performance against the 18S rDNA sequence. Regression was onto full length 18S rDNA pairwise distances.

Hyper-variable region	Regression value (R2) on all 18S diatom.	Position in bp (with *Phaeodactylum tricornutum* as reference)	Number of sequences	Min. Percentage of identity to differentiate two species
V1	0.087	3 - 103	1664	84%
V2	0.250	92 - 290	1853	86%
V3	0.071	401 - 561	2204	86%
V4	0.070	588 - 781	2799	74%
V5	0.137	1003 - 1154	2815	87%
V7	0.264	1302 - 1565	2756	80%
V8	0.256	1448 - 1565	2501	84%
V9	0.219	1640 - 1755	1777	70%
18S	-	1 - 1767	2947	93%

Elucidating reference database

Most of the 18S rDNA (region encoding the 18S rRNA) is highly conserved and is generally used for phylogenetic studies at higher taxonomic levels. PR2 reference database contains 2947 full length 18S diatom sequences from 718 diatom species (Guillou et al., 2013). From this resource, a V9 reference database comprising V9 hypervariable region constituting 361 unique diatom species was obtained. These 718 unique reference sequence were aligned and identical (100% identity) were dropped. This gave 361 unique reference sequences for further analysis.

Defining Hypervariable regions (V1-V9) in diatom sequences

Sequence variation along the entire length of 2947 18S rDNA sequences was quantified in terms of Shannon entropy and used to define hypervariable regions (V1-V9; **Figure D2**). Secondary structure prediction and the hyper-variable regions identification were done using *RNAstructure* program (Reute and Mathews, 2010). This secondary structure prediction was done on phaeodactylum sequence to identify Hypervariable region (they form specific loop structure). Using this sequences, we extracted hypervariable regions with *V-xtractor* (Hartmann et al., 2010).

Performance of V1-V9 region using phylogeny and pairwise distances

Further, to analyze the performance of V1-V9 region we used p-distance between the sequences as a measure. Pairwise distances were calculated for 2497 diatom 18S rDNA sequences using Kimura-2-parameter model. Length variation and genetic distances were shown (**Table D1**, **Figure D3**). As suggested in other eukaryotic species, a combination of hypervariable sites could help in better discrimination of different species of diatoms. For diatoms, V4 region is shown to be of higher length and showed better performance in terms of diversity determination. Regression of V1-V9 p-distance by NJ on to that of 18S sequence shows that V5 could better explain the phylogeny, followed by V4. Although the mean genetic distances were better in V4 and V9, they may not explain the phylogeny well. V9 performance was less than that of 18S. Taxa assignment at less than 70 % identity in V9 region is not recommended for diatoms.

Differentiating 4 phylogenetic groups of Diatoms

Four prominent phylogenetic clades of diatoms such as Radial centric, polar centric, Araphid pennate, raphid pennate were already known. V4 and V9 sequences were used to check the performance in differentiating the four groups. Each of the hypervariable regions and full-length 18S rDNA sequences were aligned using MUSCLE and phylogenetic inference was done with NJ algorithm using pairwise distances in MEGA5 (Tamura et al., 2011). The tree was statistically tested using 1000 bootstrap.V9 could not differentiate the araphid and raphid pennates, as they were placing them in the same branch of the phylogenetic tree (**Figure D4**).

Diatom v9 reference and the diversity

718 species are available and of those 519 have V9 regions, but UCLUST based non-redundant removal (at 100% identity) resulted in 361 V9 sequences. At least 158 species have identical V9 region. Thus, approximately 30.4% species could not be detected by V9 reference sequences and remained ambiguous.

V9 distance Vs Total SSU rDNA distance

V9 region may not be compared with the full-length 18S sequences in terms of genetic distances. The slope of the regression line between V9 and 18S is shown in the graph. V9 and SSU were poorly correlated with R value of 0.46 (**Figure D5**).

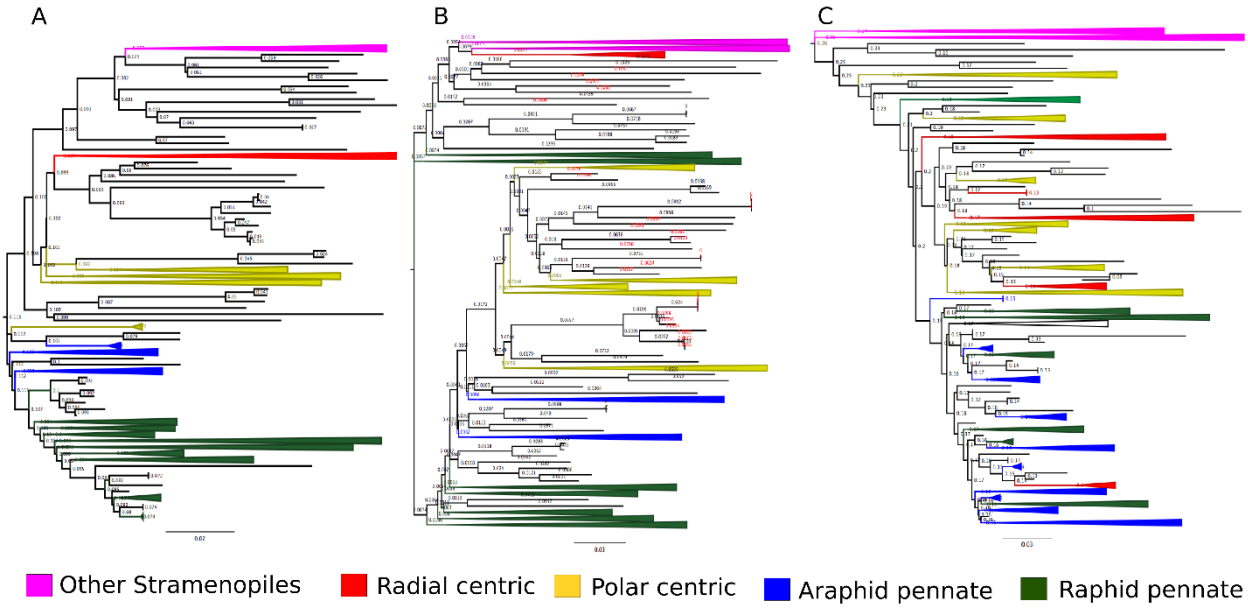


Figure D4. Phylogenetic inference from the 18S rDNA sequence of Bacillariophyta. A. Full-length 18S rDNA sequence phylogeny. B. V4 sequence phylogeny. C. V9 region rDNA sequence based phylogeny. The color codes for branches are indicated under the figure.

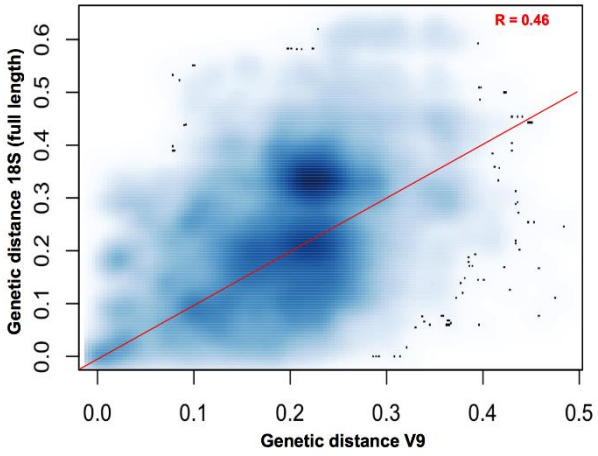


Figure D5. Comparison of complete 18S and V9 region of small-subunit rDNA sequences in establishing the genetic distances of Bacillariophytes.

E. Supplementary Information to Chapter 2

Table E1. Abundance, richness, Shannon diversity indices and dominant genera for each station.

Stations	Abundance		Richness		Shannon		Dominant Diatom Genus
	Value	Rank	Value	Rank	Value	Rank	
Mediterranean Sea:							
st4	40748	7	37	6	2.41591	1	Attheya [29%]
st7	91572	3	40	3	1.57674	11	Corethron [44%] Leptocylindrus[32%]
st9	62864	4	32	9	2.02098	3	Leptocylindrus [32%] Planktoniella [22%]
st11	633129	1	35	8	0.28139	13	Corethron [95%]
st16	37527	8	40	4	2.28414	2	Planktoniella [29%]
st18	22354	9	37	7	1.95089	5	Pseudo-nitzschia [38%]
st20	7937	12	23	13	1.31889	12	Pseudo-nitzschia [58%]
st22	100792	2	48	1	1.92326	6	Chaetoceros [26%] Pseudo-nitzschia [23%]
st23	46308	6	41	2	1.95124	4	Chaetoceros [46%]
st24	14687	10	30	11	1.78773	7	Pseudo-nitzschia [34%] Chaetoceros [29%]
st25	50267	5	38	5	1.67301	10	Pseudo-nitzschia [48%]
st26	14598	11	31	10	1.75476	9	Pseudo-nitzschia [44%]
st30	4499	13	28	12	1.7745	8	Thalassiosira [53%]
Red Sea:							
st31	13511	3	30	4	1.35253	4	Haslea [67%]
st32	11437	4	36	2	2.53472	2	Chaetoceros [21%] Thalassiosira [18%]
st33	34521	1	35	3	1.68706	3	Haslea [59%]
st34	19177	2	38	1	2.83116	1	Proboscia [12%] Actinocyclus [12%] Leptocylindrus [11%]
Indian Ocean:							
st36	97274	2	40	4	1.56619	8	Pseudo-nitzschia [62%]
st38	24843	6	33	7	2.24396	4	Chaetoceros [34%]
st39	13195	8	27	9	1.75788	6	Planktoniella [49%] Pseudo-nitzschia [21%]
st41	19383	7	38	6	2.49494	3	Leptocylindrus [28%] Chaetoceros [13%]
st42	39651	5	45	3	2.11323	5	Fragilariopsis [30%] Chaetoceros [19%] Leptocylindrus [19%]
st44	2316	11	22	11	0.85837	11	Leptocylindrus [84%]
st45	4401	9	28	8	0.98045	9	Leptocylindrus [75%]
st48	4072	10	23	10	0.96886	10	Leptocylindrus [76%]
st52	105006	1	39	5	1.5762	7	Proboscia [40%]
st64	97233	3	45	1	2.78411	1	Proboscia [12%] Chaetoceros [12%] Eucampia [10%] Actinocyclus [9%]
st65	95349	4	45	2	2.71004	2	Chaetoceros [16%] Eucampia [14%] Actinocyclus [12%] Attheya [12%]
Atlantic Ocean:							
st66	100418	3	39	4	1.96222	6	Actinocyclus [29%] Leptocylindrus [26%]
st67	284672	2	35	7	1.43888	8	Actinocyclus [61%]
st68	20812	5	40	2	2.13117	3	Actinocyclus [33%]
st70	29524	4	38	5	1.47547	7	Actinocyclus [61%]
st72	9480	8	35	8	2.15236	2	Leptocylindrus [34%] Chaetoceros [21%]
st76	15675	6	37	6	2.01959	4	Planktoniella [44%]
st78	10520	7	43	1	2.57093	1	Actinocyclus [22%] Planktoniella [18%]
st82	337579	1	39	3	1.97312	5	Chaetoceros [36%] Actinocyclus [21%]
Antarctic Ocean:							
st84	806263	2	35	2	1.17128	2	Thalassiosira [67%]
st85	2181143	1	43	1	1.17829	1	Chaetoceros [54%] Fragilariopsis [32%]
Pacific Ocean:							
st98	6493	9	37	9	2.30074	1	Actinocyclus [26%] Planktoniella [16%]
st100	411914	1	46	4	1.0325	8	Leptocylindrus [52%] Actinocyclus [39%]
st102	268658	2	45	5	1.5376	6	Actinocyclus [50%] Leptocylindrus [28%]
st109	40942	8	49	2	1.83753	3	Proboscia [47%]
st111	45577	7	40	8	1.63998	5	Leptocylindrus [40%] Actinocyclus [25%] Proboscia [22%]
st122	242919	3	48	3	1.6693	4	Actinocyclus [49%]
st123	163881	4	49	1	1.88277	2	Attheya [32]
st124	96529	5	41	7	1.48021	7	Attheya [60]
st125	80156	6	44	6	0.97718	9	Actinocyclus [74%]

F. Supplementary Information to Chapter 3

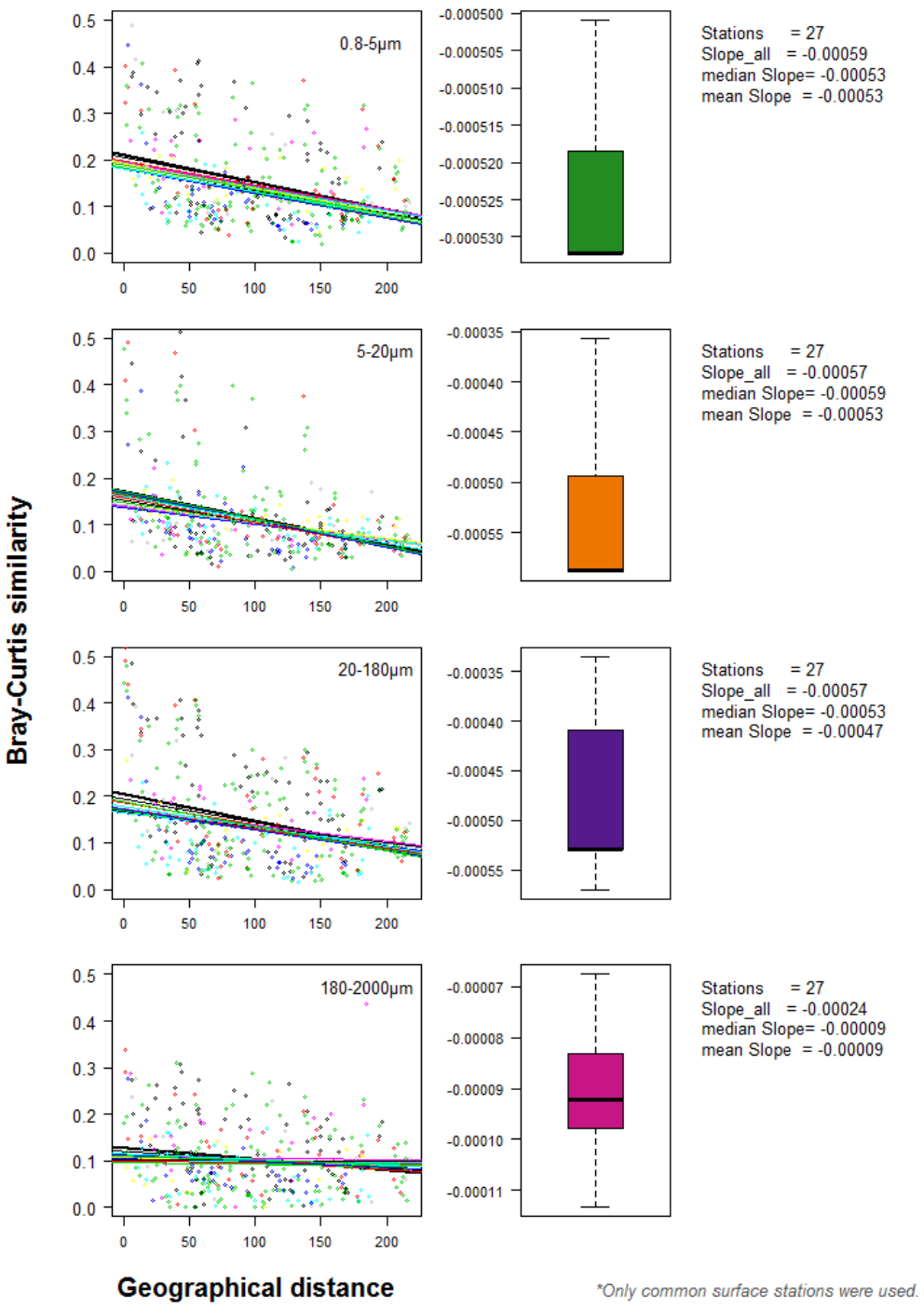


Figure F1. Estimating bias and variance using Jackknife method. This method works by calculating the statistic of interest leaving out one sample at a time.

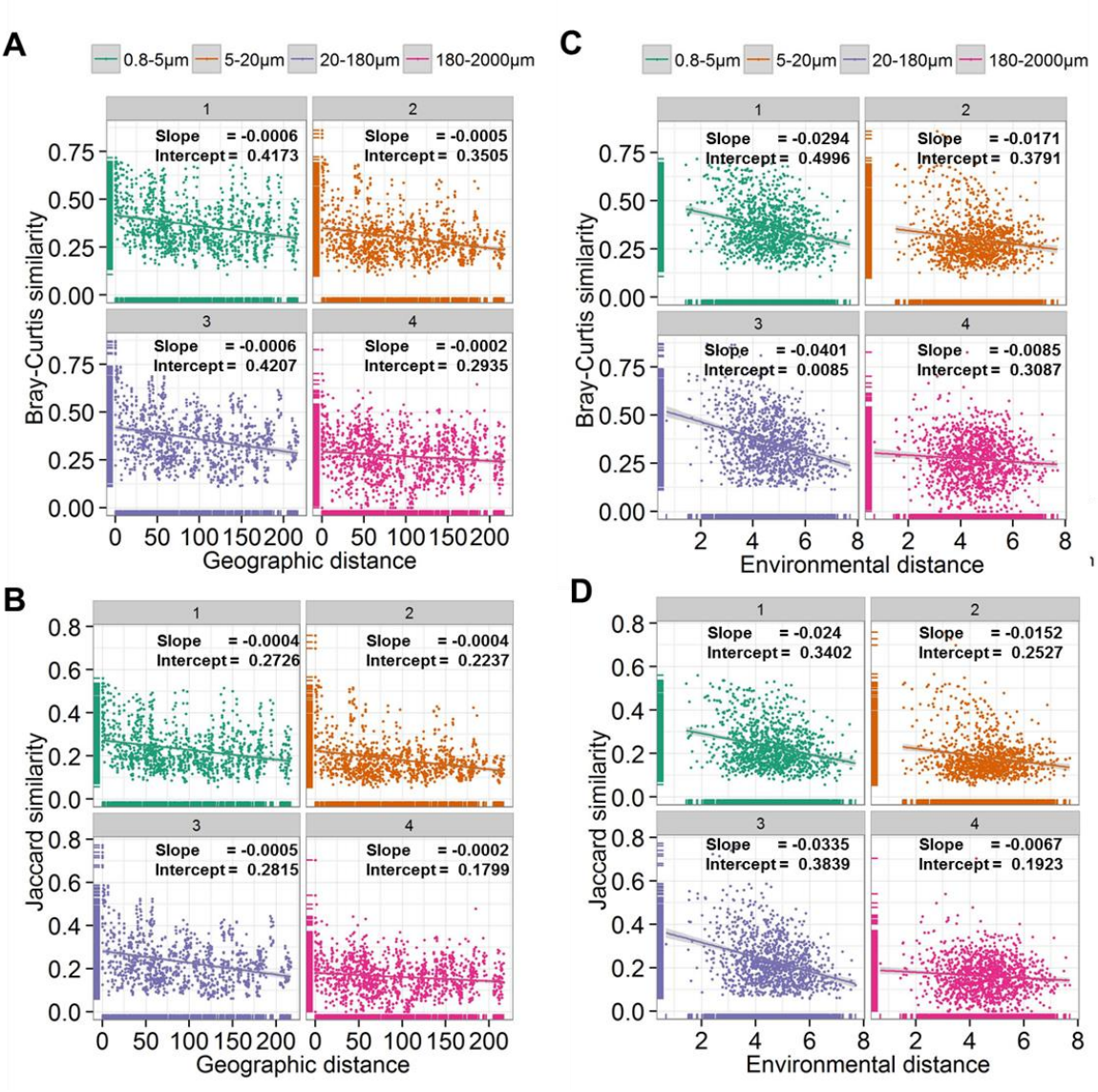


Figure F2. Relationship between community similarity using Swarms and geographic distances/ environmental distances. (A) Bray-Curtis and **(B)** Jaccard similarity of diatom communities plotted against geographic distances between sites. **(C)** Bray-Curtis, and **(D)** Jaccard similarity of diatom communities plotted against environmental distances between sites.

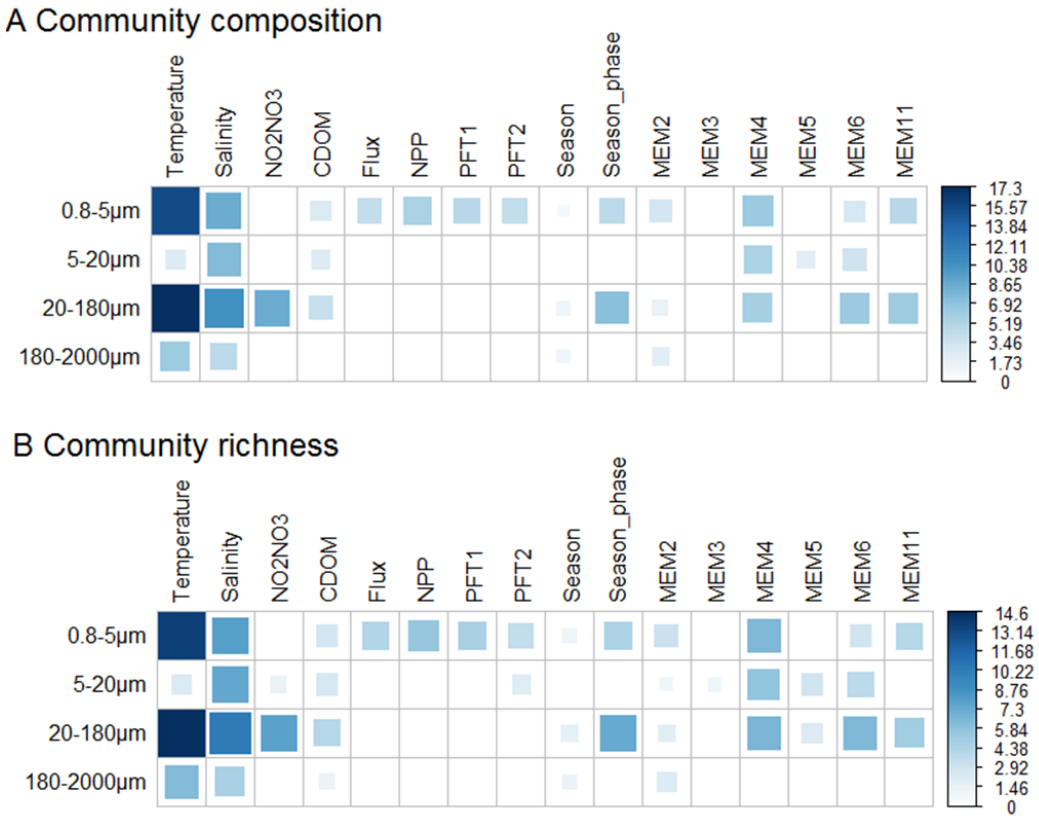


Figure F3. Permutation-based multiple regression on distance matrices (MRM) using Swarms. Environmental variables significantly contributing to the variation in diatom community similarity are shown. Each environmental variable was used as an independent matrix. R-squared regression coefficient of each environmental variable is expressed as percent.

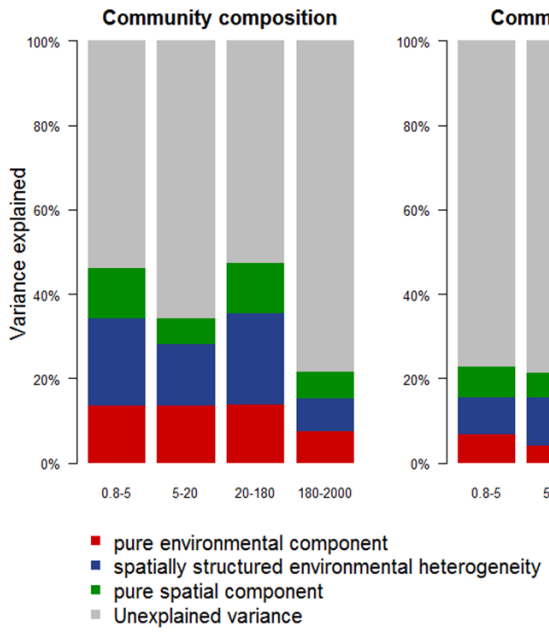


Figure F4. Variation in community composition and richness explained by environmental and spatial variables and their shared effects computed using Swarms.

G. Supplementary Information to Chapter 4

Table G4.1. Summary of all the environmental variables across the 46 sampling stations.

Descriptor	min	max	mean	median	corr	pval
Salinity	33.37	40.20	36.22	35.88	-0.37	0.000
Nitrates	-2.20	24.42	3.06	0.89	0.47	0.000
Chloro.HPLC	0.00	1.02	0.23	0.20	0.27	0.000
Angular.scattering.coef	0.00	0.00	0.00	0.00	0.40	0.000
Part.backscattering	0.00	0.00	0.00	0.00	0.42	0.000
Fcdom	0.81	2.82	1.62	1.61	0.25	0.000
beam.attenuation.coef	0.02	0.32	0.08	0.06	0.30	0.000
Flux_150m	0.01	12.55	2.10	0.94	0.37	0.000
Flux_400m	0.02	8.85	1.46	0.80	0.35	0.000
Depth.Max.Fluo	21.00	175.83	63.69	54.33	-0.25	0.000
NO2	0.00	1.47	0.15	0.05	0.35	0.000
PO4	0.00	1.94	0.41	0.27	0.39	0.000
NO2NO3	0.00	28.64	3.58	0.56	0.39	0.000
SI	0.32	16.55	2.36	1.47	0.33	0.000
NPP_month	75.19	1455.03	448.44	420.55	0.28	0.000
OG.Shannon	7.04	7.32	7.15	7.15	0.23	0.000
nbr_zoo	68.28	2806.25	718.47	477.05	0.36	0.000
biovolume_zoo	6.46	866.58	190.04	106.16	0.32	0.000
Long	-142.58	73.90	-20.97	4.69	-0.21	0.001
Mean.Depth.Nitrocline	41.14	450.20	194.96	149.00	-0.22	0.001
NPP_8d	63.56	1428.93	442.42	422.49	0.23	0.001
Lat	-60.23	42.21	-4.18	-8.90	-0.21	0.002
Lyapunov_exp	0.00	0.32	0.04	0.03	0.20	0.002
total.pico.euk	0.00	19090.56	4985.28	3912.88	0.21	0.002
OG.Evenness	0.73	0.77	0.74	0.74	0.19	0.006
Temp	1.84	30.50	22.56	24.70	-0.18	0.009
bbp470	0.00	0.00	0.00	0.00	0.17	0.011
Okubo.Weiss	-0.86	1.83	0.03	0.00	0.15	0.022
total.auto	0.00	403151.23	128991.83	123011.07	-0.14	0.036
Gene.Simpson	1.00	1.00	1.00	1.00	0.14	0.043
Gene.Inv.Simpson	37065.90	481037.50	166887.76	149145.00	0.13	0.054
total.hetero	10638.50	2216958.86	590853.71	480515.42	0.13	0.057
SEASON.PART	1.00	3.00	1.82	2.00	0.11	0.095
miTAG.SILVA.Shannon	6.16	6.86	6.55	6.56	0.11	0.104
grad_SST_adv	0.14	2.22	0.94	0.76	0.11	0.110
FLAG.SEASON.S.Su.A.W	1.00	4.00	2.81	3.00	0.10	0.136
pente_Zoo	-1.07	-0.60	-0.79	-0.79	0.08	0.232
miTAG.SILVA.Richness	1998.18	3993.38	2644.96	2544.51	0.08	0.236
Oxygen	103.92	338.30	200.65	200.23	0.08	0.255
MLD	5.00	187.25	47.13	34.57	0.07	0.313
miTAG.SILVA.Chao	2709.60	6141.60	3812.91	3734.95	0.07	0.333
OG.Richness	12175.00	21731.60	15622.96	15374.40	-0.06	0.365
miTAG.SILVA.ace	2679.20	6033.80	3706.58	3634.61	0.06	0.369
Gene.Eveness	0.85	0.93	0.90	0.90	0.06	0.412
X16S.18S.ratio	2.95	88.12	29.38	25.25	-0.05	0.424
Pressure	5.50	185.50	39.18	6.00	0.05	0.446
retention	0.00	69.36	11.15	2.92	-0.05	0.447
AMODIS.PARm.Einsteins	10.26	48.73	35.85	39.62	-0.05	0.461
Gene.Shannon	11.64	14.31	13.47	13.49	0.04	0.542
Depth.Max.N2	23.00	199.62	86.89	79.90	-0.04	0.563
AMODIS.PAR8d.Einsteins	7.74	51.72	34.38	37.50	-0.02	0.773
Depth.Max.O2	28.00	608.50	85.24	31.48	-0.02	0.809
md.Mean_Flux_400m.md.Mean_Flux_150m	0.06	3.25	1.07	0.76	0.01	0.875
Gene.Chao1	1141225.00	7438605.00	4686979.70	4598550.00	0.01	0.910
Depth.Min.O2	111.33	869.00	375.36	374.00	0.01	0.918
abs.md.Mean_Lat.	0.00	60.23	20.86	18.80	-0.01	0.918
total.bacteria	10638.50	2448660.64	683277.22	620056.21	0.01	0.935
Gene.Richness	847851.00	5118407.00	3339284.15	3283653.50	0.00	0.998

Table G4.2. Taxonomic Composition of each cluster. (A) Richness. (B) Abundance.

Taxonomic Details	Clusters									Grand Total
	I	II	III	IV	V	VI	VII	VIII	IX	
AP					38			1	1	40
AP|Asterionellopsis|Asterionellopsis+glacialis	1				1					2
AP|Catacombas|Catacombas+gaillonii					1					1
AP|Rhaphoneis|Rhaphoneis+sp.					1					1
AP|Synedra|Synedra+sp.					1					1
AP|Synedra|Synedra+toxoneides					3				2	5
PC		76		6	89	2		2	19	194
PC|Attheya|Attheya+longicornis					3				12	15
PC|Chaetoceros					3	1				4
PC|Chaetoceros|Chaetoceros+muellerii					3					3
PC|Chaetoceros|Chaetoceros+radicans					1					1
PC|Chaetoceros|Chaetoceros+rostratus	1		3		100			2	2	108
PC|Chaetoceros|Chaetoceros+sp.	6			50	14			24	1	95
PC|Cyclotella|Cyclotella+choctawhatcheeana									1	1
PC|Ditylum|Ditylum+brightwellii					1					1
PC|Eucampia	1				8				1	10
PC|Minidiscus					1					1
PC|Minutocellus					5					5
PC|Odontella|Odontella+sinensis						1				1
PC|Pierrecomperia|Pierrecomperia+catenuloides					1					1
PC|Planktoniella|Planktoniella+sol									7	7
PC|Porosira						1				1
PC|Skeletonema					3					3
PC|Skeletonema|Skeletonema+grevillei	2				1					3
PC|Skeletonema|Skeletonema+menzellii	1				1					2
PC|Thalassiosira		1		1	22	1	1		45	71
PC|Thalassiosira|Thalassiosira+aestivalis									1	1
PC|Thalassiosira|Thalassiosira+concaviuscula				1						1
PC|Thalassiosira|Thalassiosira+delicatula					1					1
PC|Thalassiosira|Thalassiosira+minima					1					1
PC|Thalassiosira|Thalassiosira+punctigera				1	1				24	26
PC|Thalassiosira|Thalassiosira+sp.					1				1	2
PC|Thalassiosira|Thalassiosira+tumida								8		8
PC|Thalassiosira|Thalassiosira+weissflogii	1							8		9
RC					8					8
RC|Actinocyclus|Actinocyclus+curvatulus			4		22		4	2	36	68
RC|Corethron					3	2				5
RC|Corethron|Corethron+hystrix			4		1		1			6
RC|Corethron|Corethron+inerme	2		39	1	1	9	5	1		58
RC|Coscinodiscus|Coscinodiscus+radiatus					1				1	2
RC|Coscinodiscus|Coscinodiscus+sp.					1					1
RC|Guinardia|Guinardia+flaccida					8				1	9
RC|Leptocylindrus				1	15				5	21
RC|Leptocylindrus|Leptocylindrus+convexus					1					1
RC|Proboscia|Proboscia+alata			1		22			8		31
RC|Rhizosolenia					2					2
RC|Rhizosolenia|Rhizosolenia+setigera								1		1
RC|Rhizosolenia|Rhizosolenia+shrubsolei				1	2					3
RC|Stellarima|Stellarima+microtrias						1				1
RP	19		5	2	77	10	39	5	68	225
RP|Bacillariophyta|Bacillariophyta+sp.					1					1
RP|Craticula|Craticula+cuspidata					1					1
RP|Cylindrotheca|Cylindrotheca+closterium					1				1	2
RP|Cymbella					1				1	2
RP|Fragilariopsis	61		1		16	6	1	1	1	87
RP|Haslea|Haslea+spicula					1				1	2
RP|Navicula	2									2
RP|Navicula|Navicula+gregaria									1	1
RP|Navicula|Navicula+radiosa					1			1		2
RP|Navicula|Navicula+salinicola					1				1	2
RP|Naviculales|Naviculales+sp.					1					1
RP|Pleurosigma|Pleurosigma+sp.					6				2	8
RP|Pseudo-nitzschia				1	13				5	19
RP|Pseudo-nitzschia|Pseudo-nitzschia+australis					2			1		3
RP|Pseudo-nitzschia|Pseudo-nitzschia+fraudulenta					1					1
RP|Pseudo-nitzschia|Pseudo-nitzschia+heimii					3				5	8
RP|Pseudo-nitzschia|Pseudo-nitzschia+pseudodelicatissima				1	2					3
RP|Pseudo-nitzschia|Pseudo-nitzschia+sp.						1				1
RP|RP_X|RP_X+sp.							2			2
unassigned	65		1	2	70	4	2	23	52	219
Grand Total	**162**	**77**	**58**	**68**	**590**	**39**	**55**	**88**	**298**	**1435**

(A)

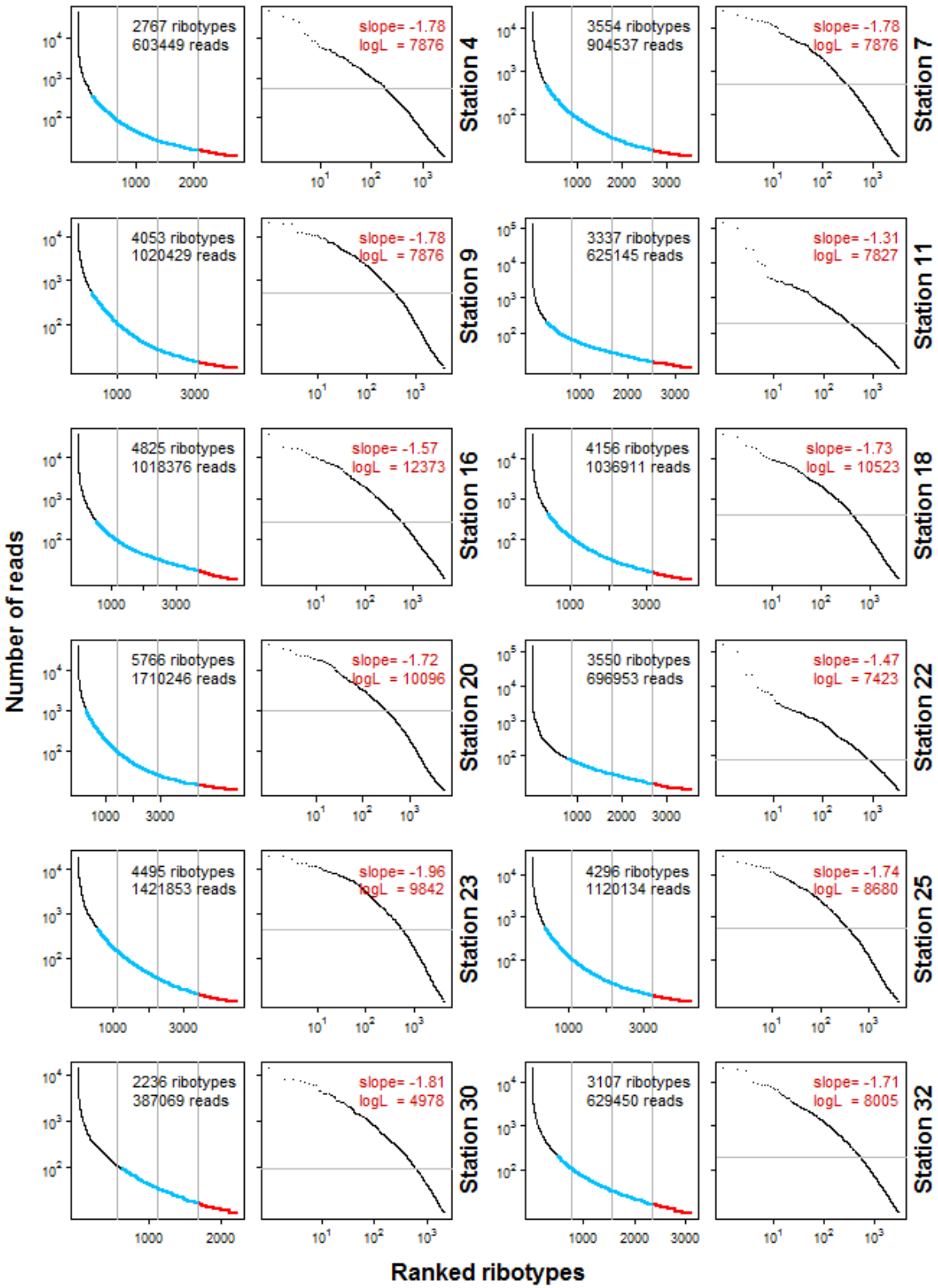
Taxonomic Details	Clusters									Grand Total
	I	II	III	IV	V	VI	VII	VIII	IX	
AP					298925			655	225	**299805**
AP|Asterionellopsis|Asterionellopsis+glacialis	111				13304					**13415**
AP|Catacombas|Catacombas+gaillonii					181					**181**
AP|Rhaphoneis|Rhaphoneis+sp.					1193					**1193**
AP|Synedra|Synedra+sp.					3662					**3662**
AP|Synedra|Synedra+toxoneides					3176				16265	**19441**
PC		584025		4376	384049	703		12432	162612	**1148197**
PC|Attheya|Attheya+longicornis					6621				81164	**87785**
PC|Chaetoceros					939	2736				**3675**
PC|Chaetoceros|Chaetoceros+muellerii					2112					**2112**
PC|Chaetoceros|Chaetoceros+radicans					212					**212**
PC|Chaetoceros|Chaetoceros+rostratus	123		629		536513			36235	2299	**575799**
PC|Chaetoceros|Chaetoceros+sp.	1200			560687	27902			234884	427	**825100**
PC|Cyclotella|Cyclotella+choctawhatcheeana									285	**285**
PC|Ditylum|Ditylum+brightwellii					434					**434**
PC|Eucampia	219				77037				10783	**88039**
PC|Minidiscus					23032					**23032**
PC|Minutocellus					14687					**14687**
PC|Odontella|Odontella+sinensis						2056				**2056**
PC|Pierrecomperia|Pierrecomperia+catenuloides					133					**133**
PC|Planktoniella|Planktoniella+sol									70852	**70852**
PC|Porosira						531				**531**
PC|Skeletonema					2166					**2166**
PC|Skeletonema|Skeletonema+grevillei	395				1350					**1745**
PC|Skeletonema|Skeletonema+menzellii	125				1551					**1676**
PC|Thalassiosira		2421		587	173793	5704	189		331270	**513964**
PC|Thalassiosira|Thalassiosira+aestivalis									1742	**1742**
PC|Thalassiosira|Thalassiosira+concaviuscula				1355						**1355**
PC|Thalassiosira|Thalassiosira+delicatula					2420					**2420**
PC|Thalassiosira|Thalassiosira+minima					180					**180**
PC|Thalassiosira|Thalassiosira+punctigera				248	212				154498	**154958**
PC|Thalassiosira|Thalassiosira+sp.					403				27668	**28071**
PC|Thalassiosira|Thalassiosira+tumida								87509		**87509**
PC|Thalassiosira|Thalassiosira+weissflogii	126							53091		**53217**
RC					86412					**86412**
RC|Actinocyclus|Actinocyclus+curvatulus			93753		135987		1132	11554	306066	**548492**
RC|Corethron					16449	257				**16706**
RC|Corethron|Corethron+hystrix			4015		9783		172			**13970**
RC|Corethron|Corethron+inerme	999		671047	262	182	2705	1315	11650		**688160**
RC|Coscinodiscus|Coscinodiscus+radiatus					12938				824	**13762**
RC|Coscinodiscus|Coscinodiscus+sp.					1739					**1739**
RC|Guinardia|Guinardia+flaccida					35020				968	**35988**
RC|Leptocylindrus				182	620267				2570	**623019**
RC|Leptocylindrus|Leptocylindrus+convexus					23490					**23490**
RC|Proboscia|Proboscia+alata			164		139596			101752		**241512**
RC|Rhizosolenia					2287					**2287**
RC|Rhizosolenia|Rhizosolenia+setigera								4508		**4508**
RC|Rhizosolenia|Rhizosolenia+shrubsolei				31366	7341					**38707**
RC|Stellarima|Stellarima+microtrias						2440				**2440**
RP	5121		1232	1058	316252	329208	313770	18294	700118	**1685053**
RP|Bacillariophyta|Bacillariophyta+sp.					708					**708**
RP|Craticula|Craticula+cuspidata					875					**875**
RP|Cylindrotheca|Cylindrotheca+closterium					1528				4455	**5983**
RP|Cymbella					1586				3489	**5075**
RP|Fragilariopsis	687342		619		12007	1918	102	1979	2565	**706532**
RP|Haslea|Haslea+spicula					339				22969	**23308**
RP|Navicula	283									**283**
RP|Navicula|Navicula+gregaria									162	**162**
RP|Navicula|Navicula+radiosa					1686			10715		**12401**
RP|Navicula|Navicula+salinicola					218				3162	**3380**
RP|Naviculales|Naviculales+sp.					9032					**9032**
RP|Pleurosigma|Pleurosigma+sp.					35405				5236	**40641**
RP|Pseudo-nitzschia				5431	121511				20131	**147073**
RP|Pseudo-nitzschia|Pseudo-nitzschia+australis					2896			1560		**4456**
RP|Pseudo-nitzschia|Pseudo-nitzschia+fraudulenta					491					**491**
RP|Pseudo-nitzschia|Pseudo-nitzschia+heimii					34311				48977	**83288**
RP|Pseudo-nitzschia|Pseudo-nitzschia+pseudodelicatissima				3775	10316					**14091**
RP|Pseudo-nitzschia|Pseudo-nitzschia+sp.						18674				**18674**
RP|RP_X|RP_X+sp.							2073			**2073**
unassigned	750916		340	3235	163332	1237	2532	242503	545098	**1709193**
Grand Total	**1446960**	**586446**	**771799**	**612562**	**3380171**	**368169**	**321285**	**829321**	**2526880**	**10843593**

(B)

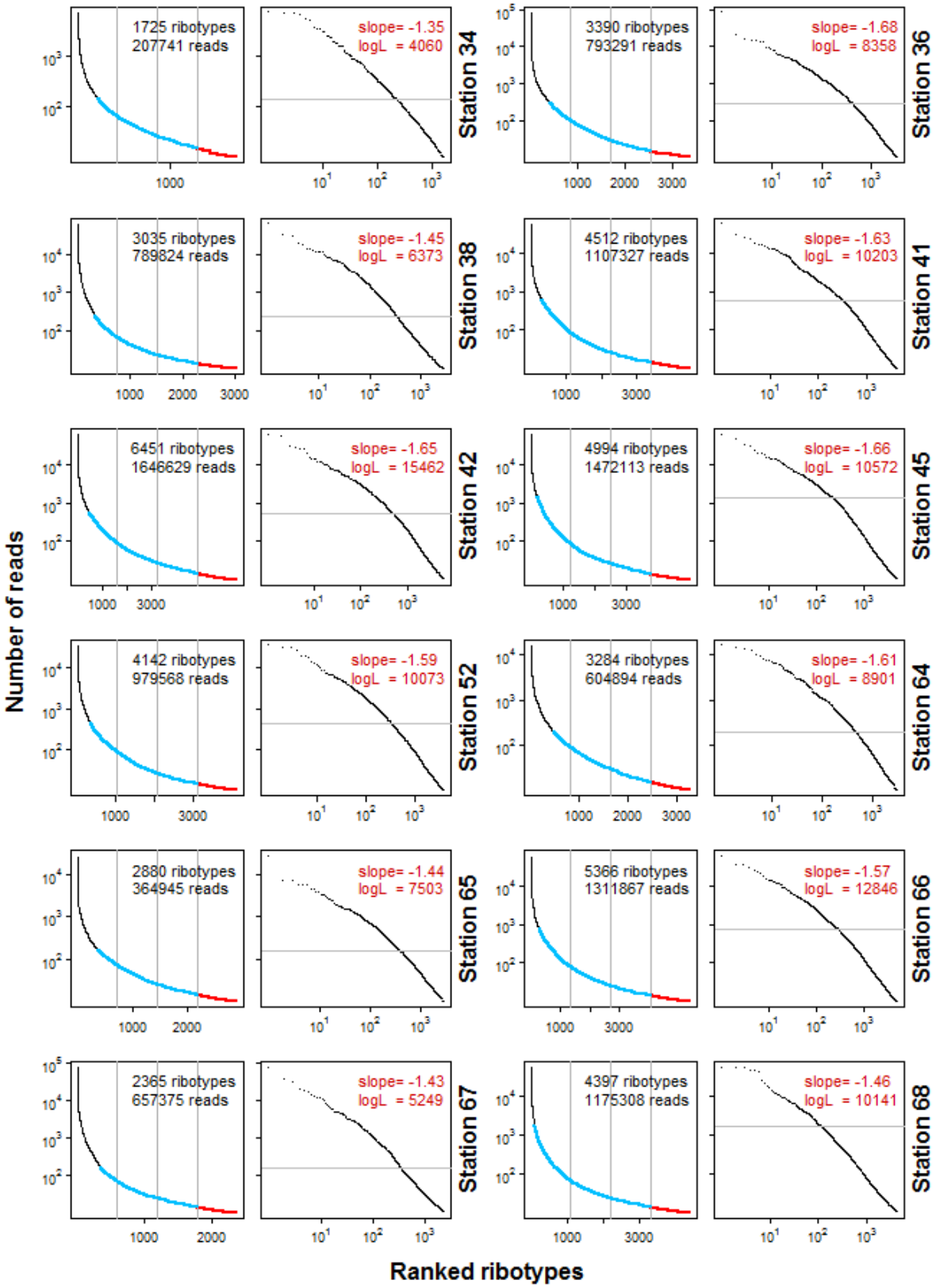
H. Supplementary Information to Chapter 5

Rank-abundance curves of empirical protistan community samples

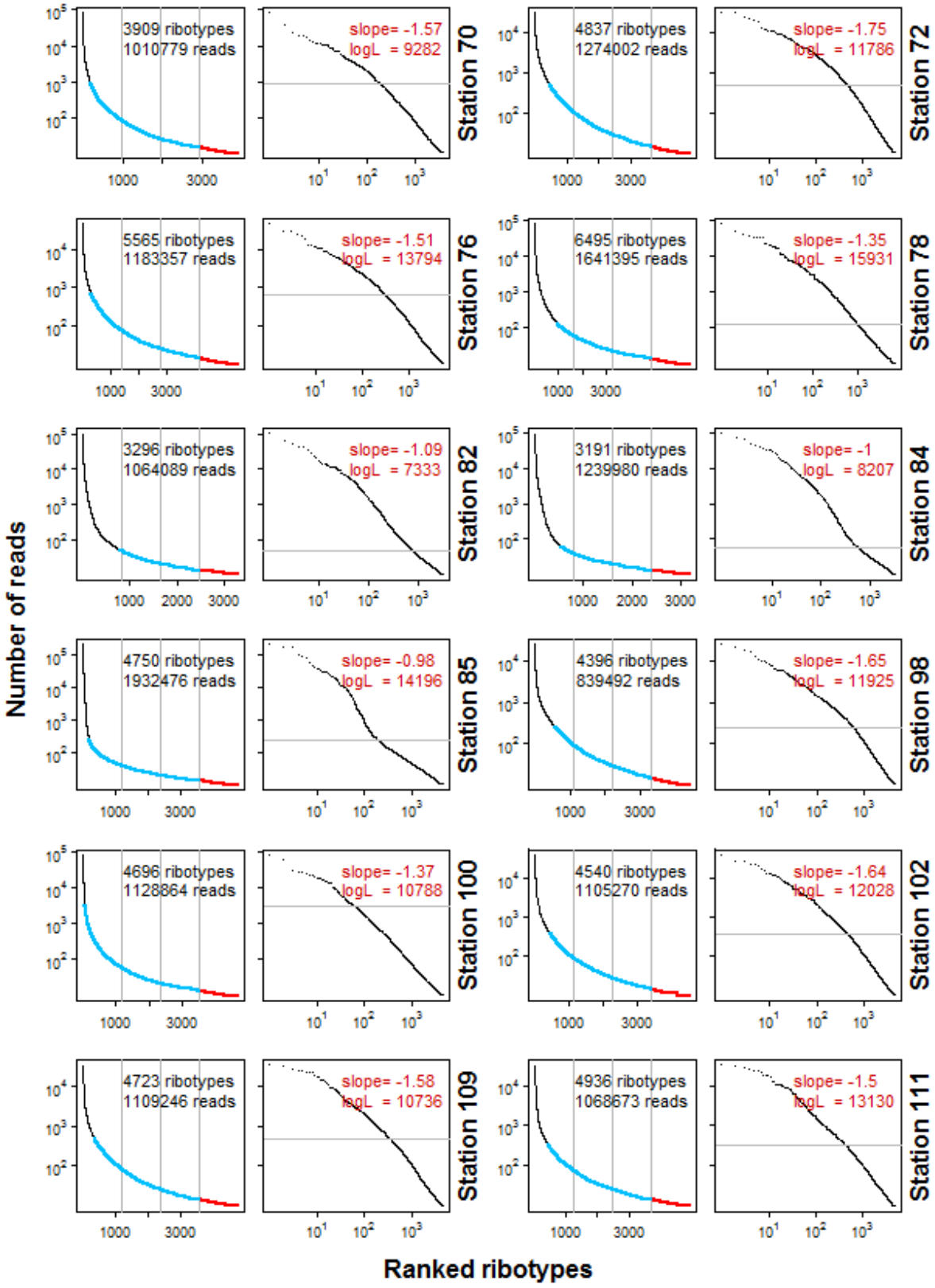
A. Pico-eukaryote (Station 4 to 32)



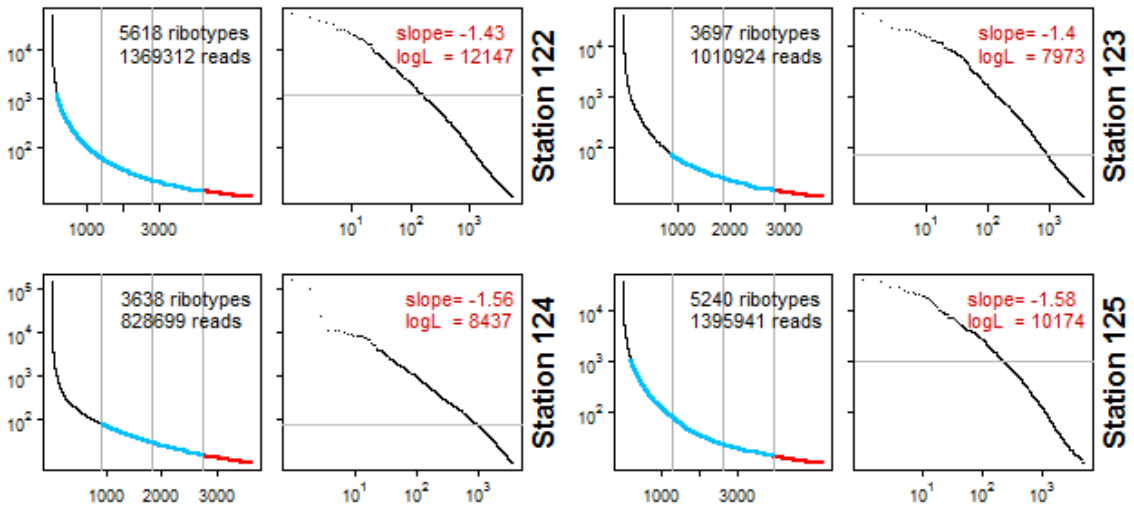
A. Pico-eukaryote (Station 34 to 68)



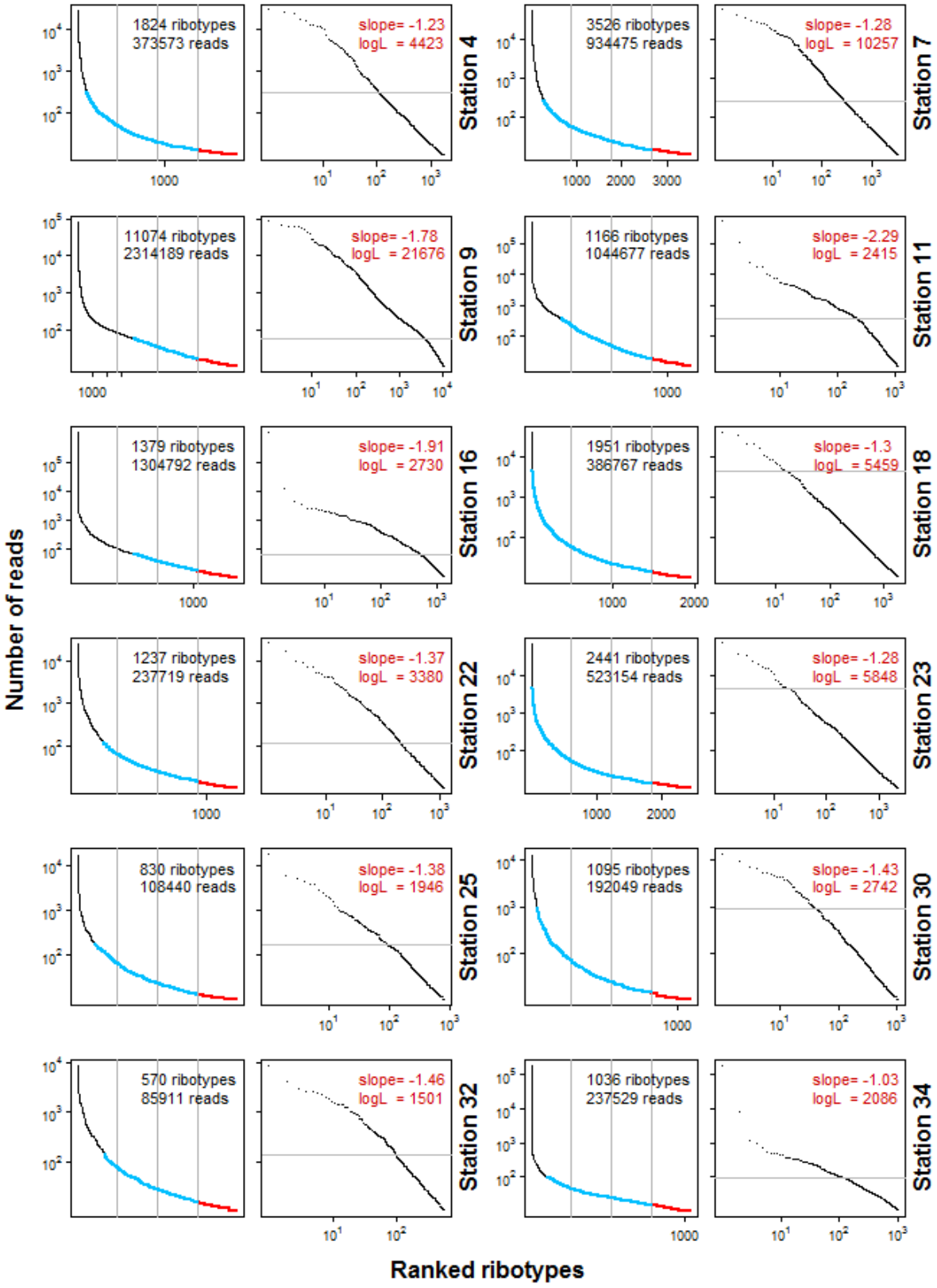
A. Pico-eukaryote (Station 70 to 111)



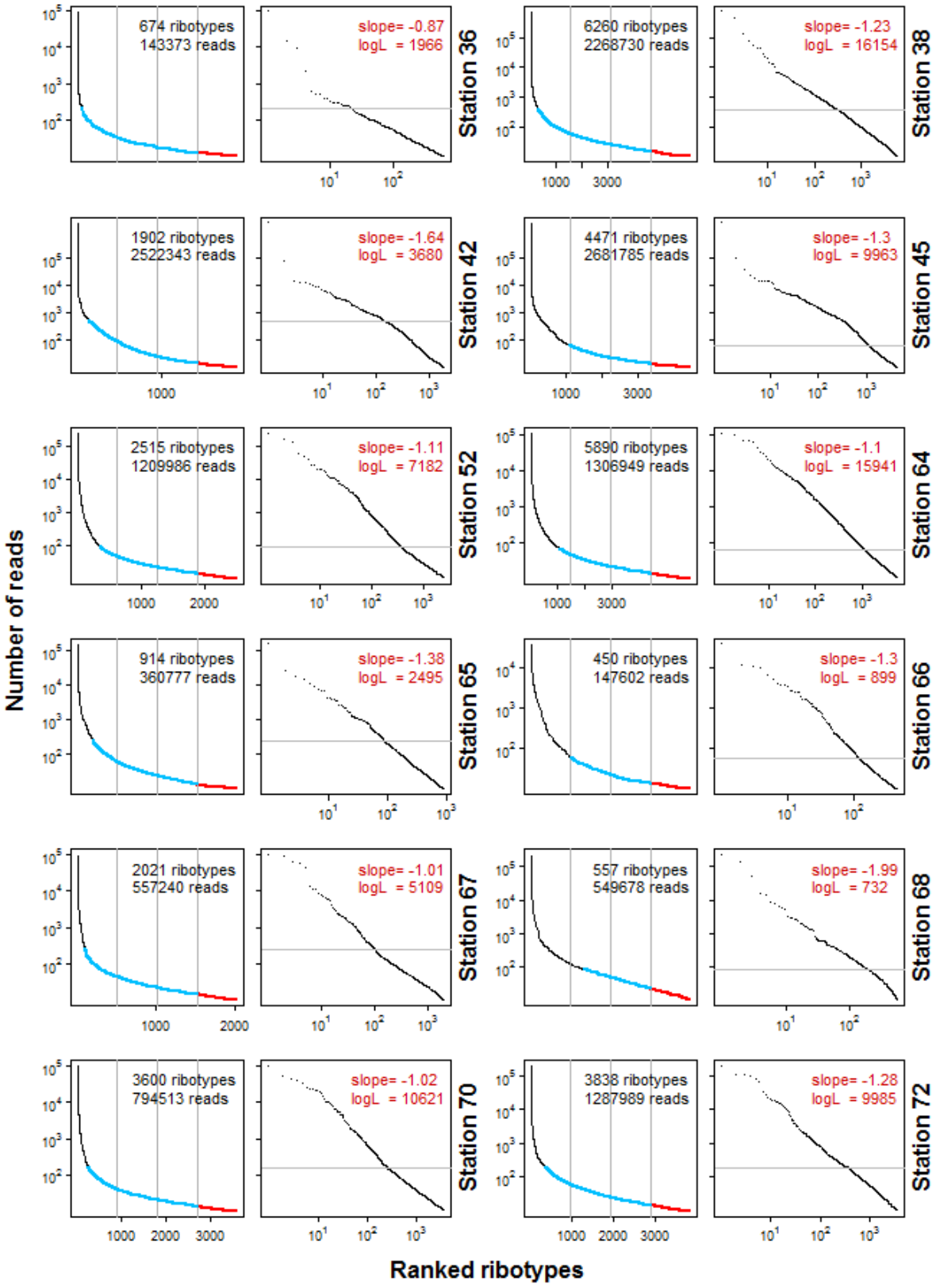
A. Pico-eukaryote (Station 122 to 125)



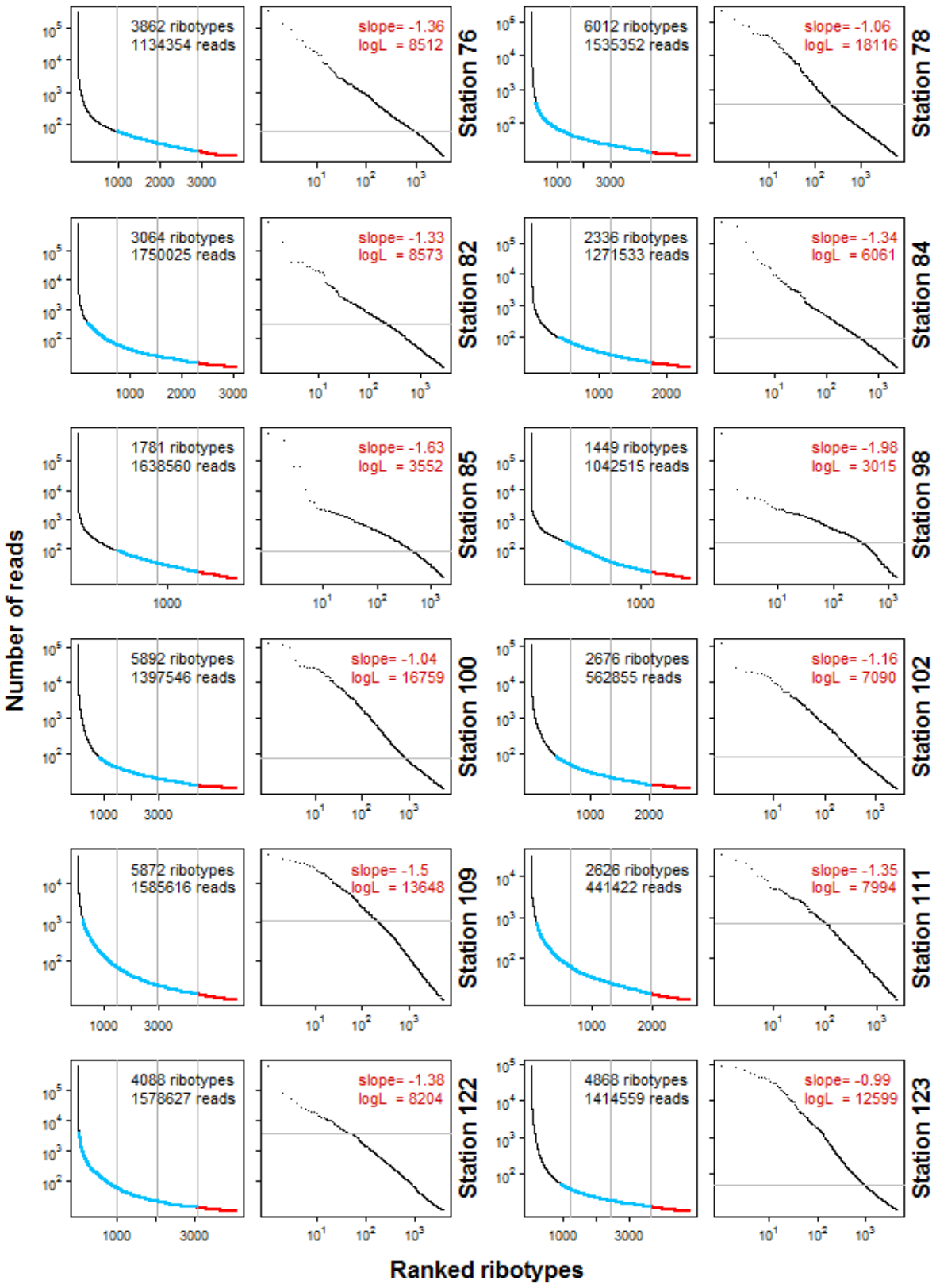
B. Nano-eukaryote (Station 4 to 34)



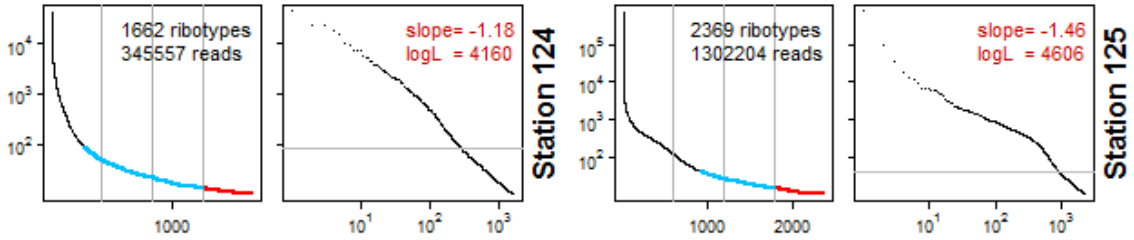
B. Nano-eukaryote (Station 36 to 72)



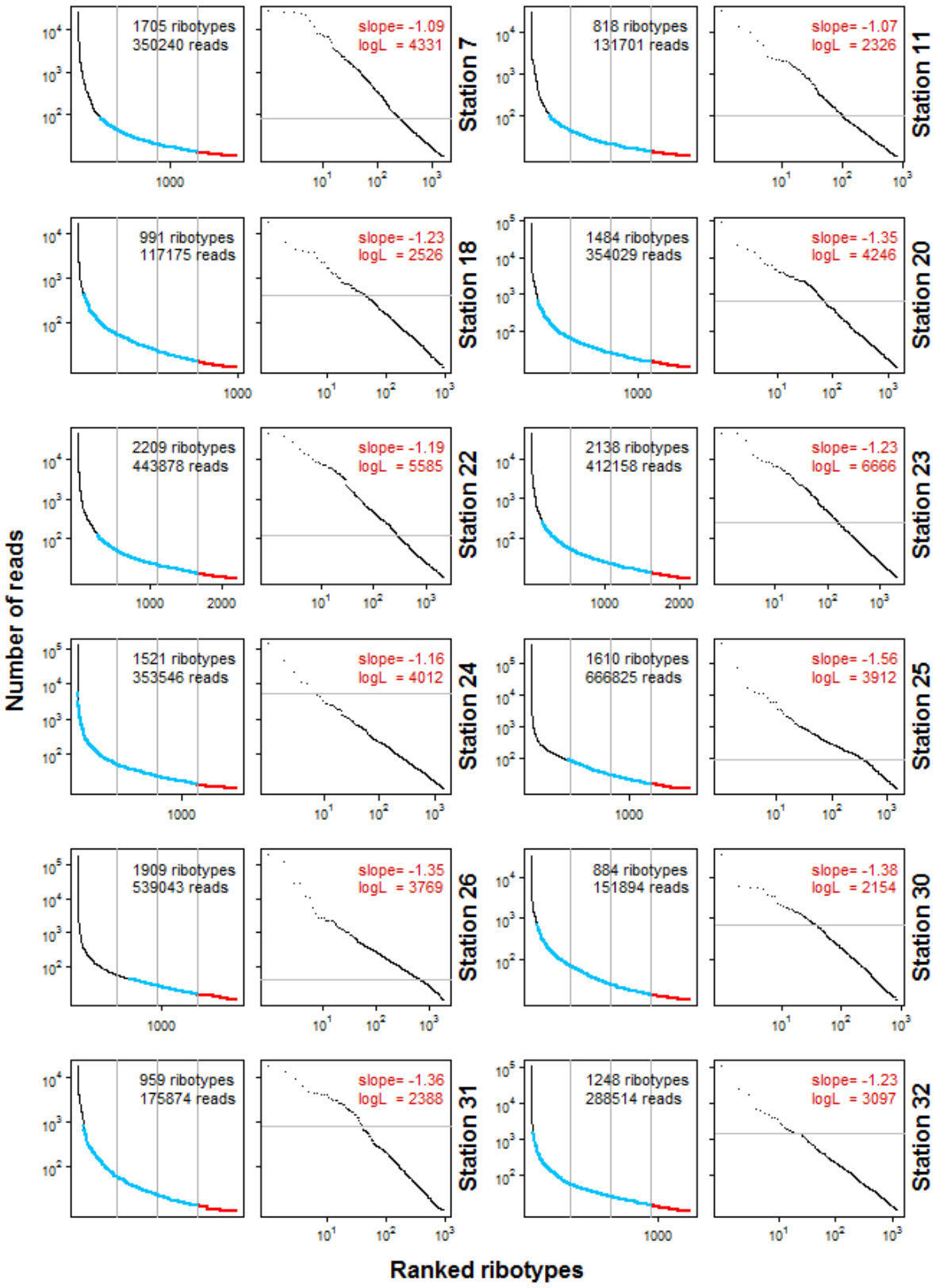
B. Nano-eukaryote (Station 76 to 123)



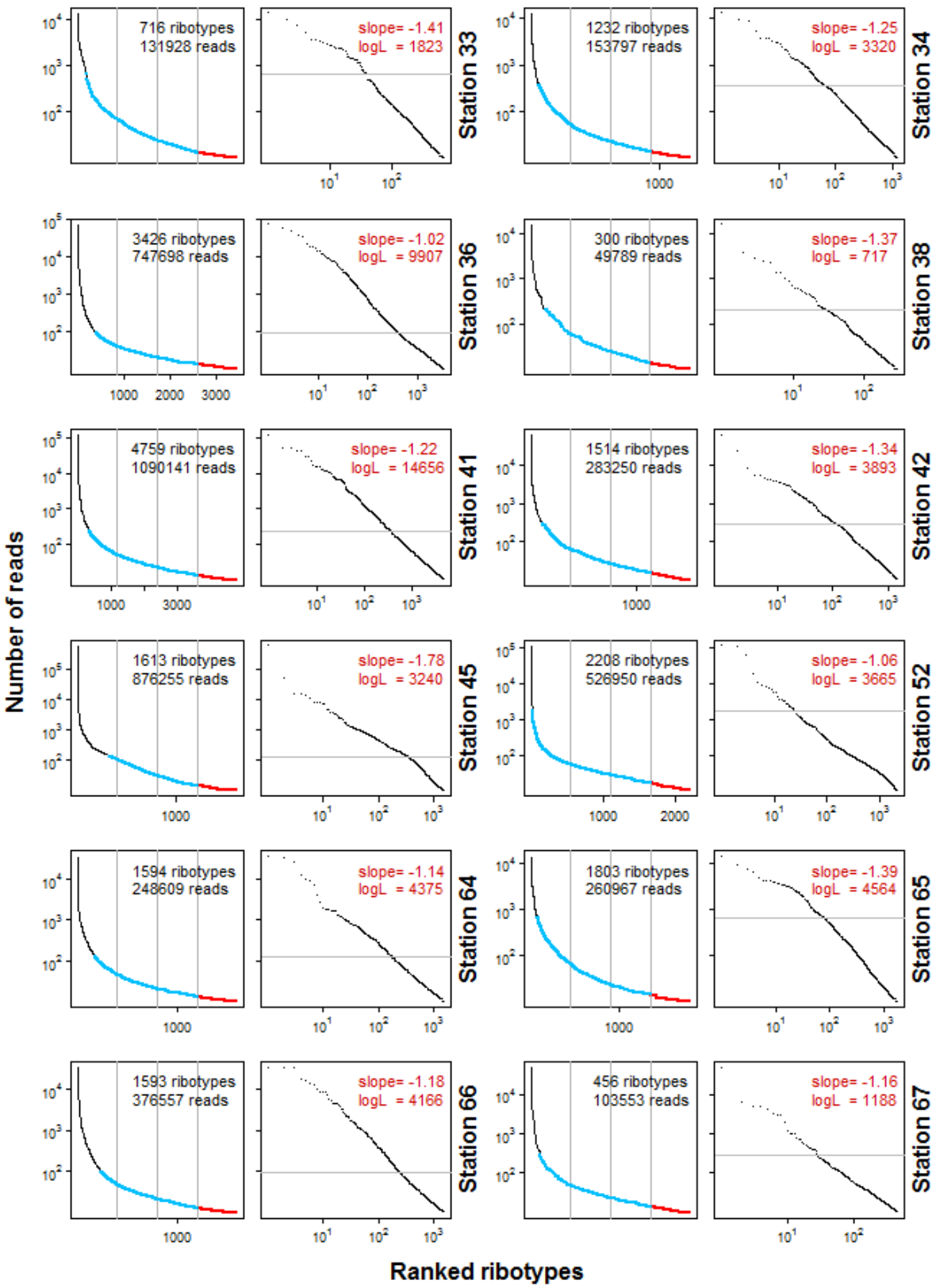
B. Nano-eukaryote (Station 124 to 125)



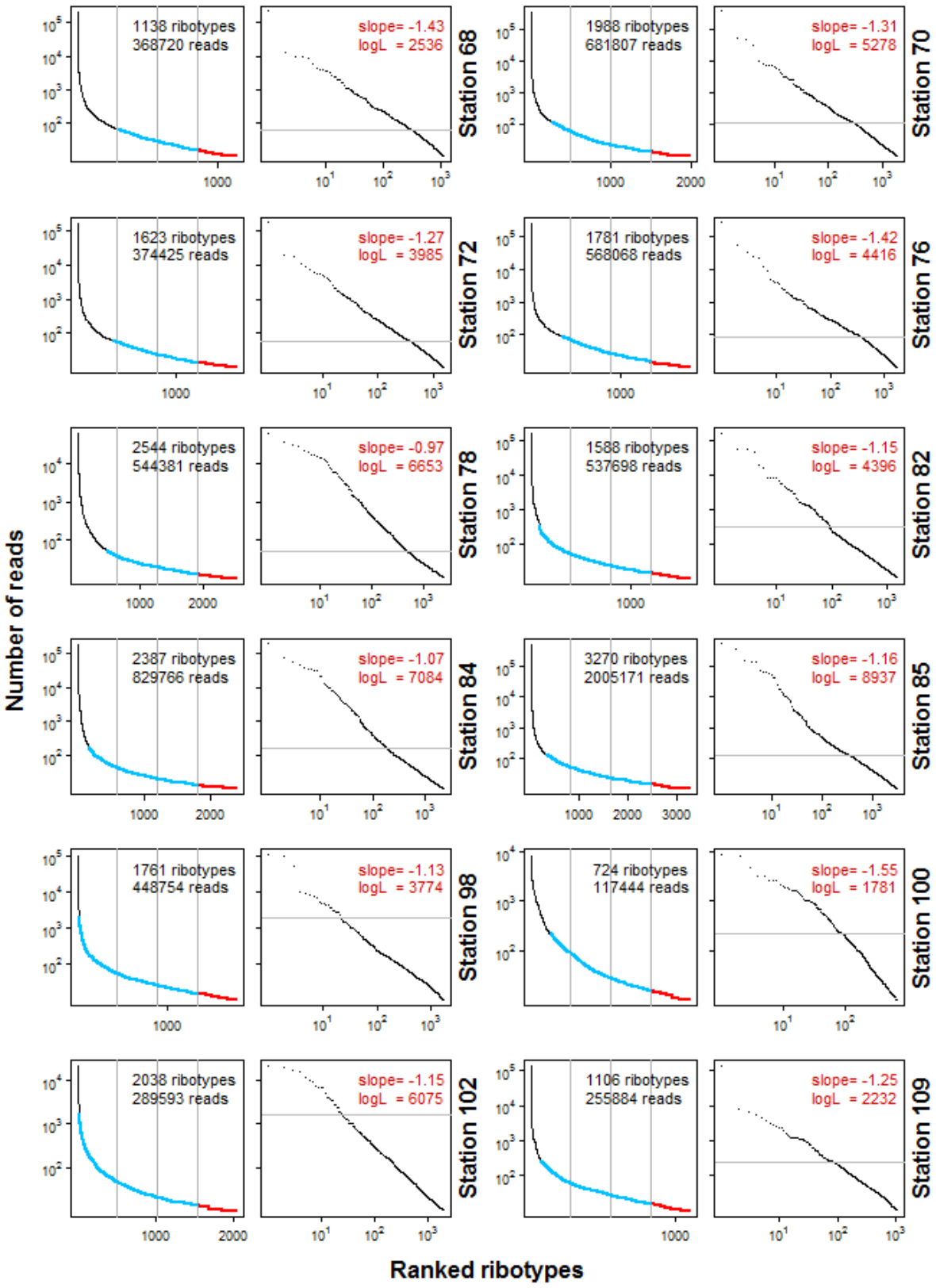
C. Micro-eukaryote (Station 4 to 32)



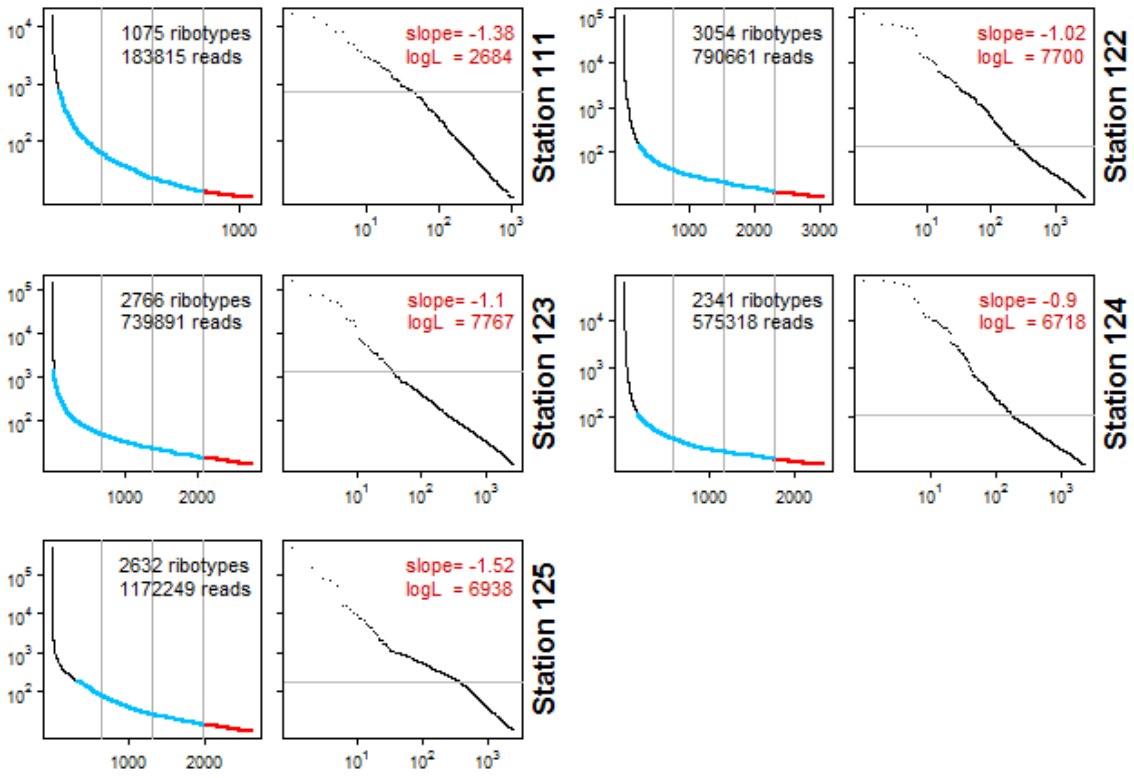
C. Micro-eukaryote (Station 33 to 67)



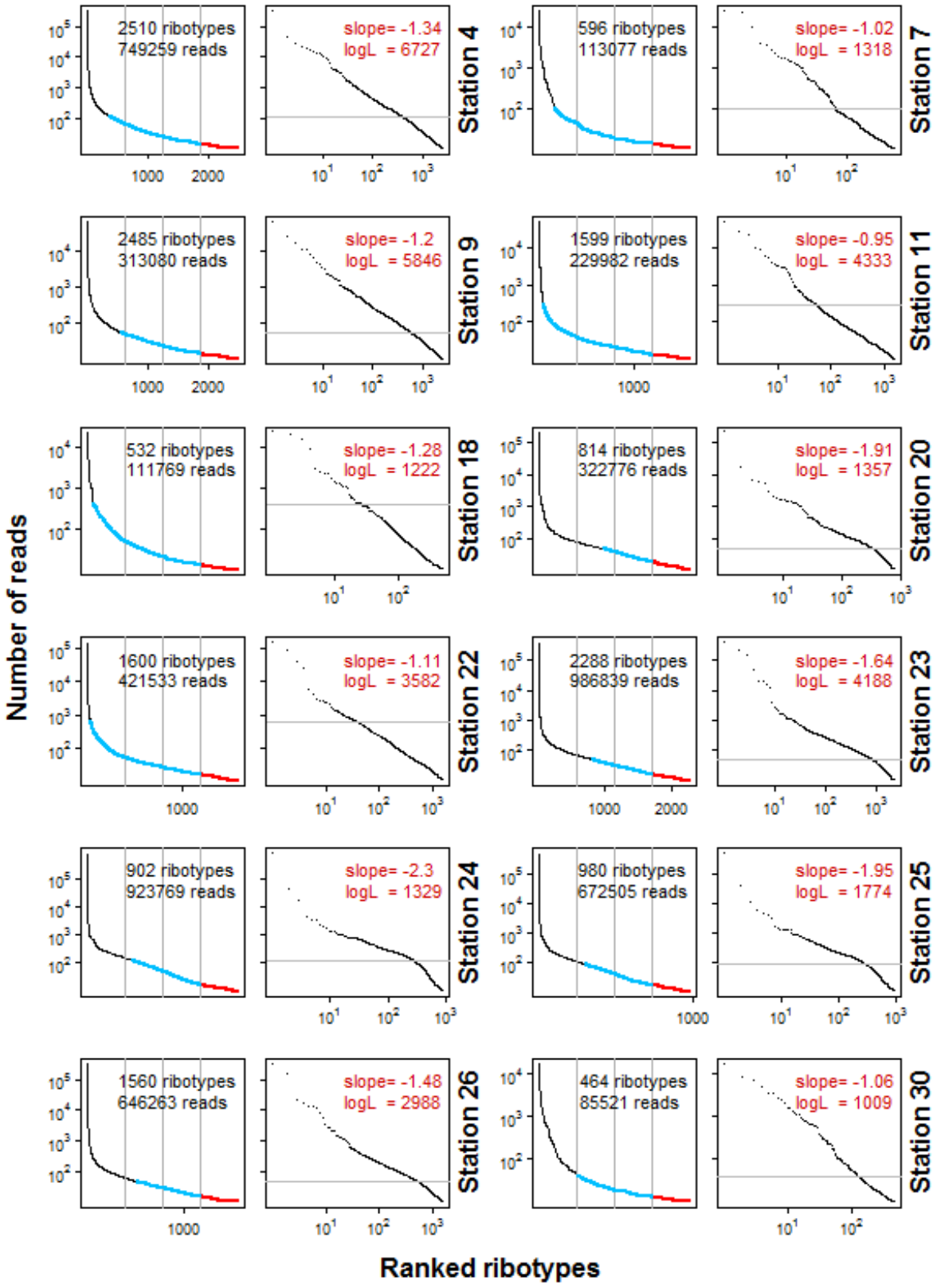
C. Micro-eukaryote (Station 68 to 109)



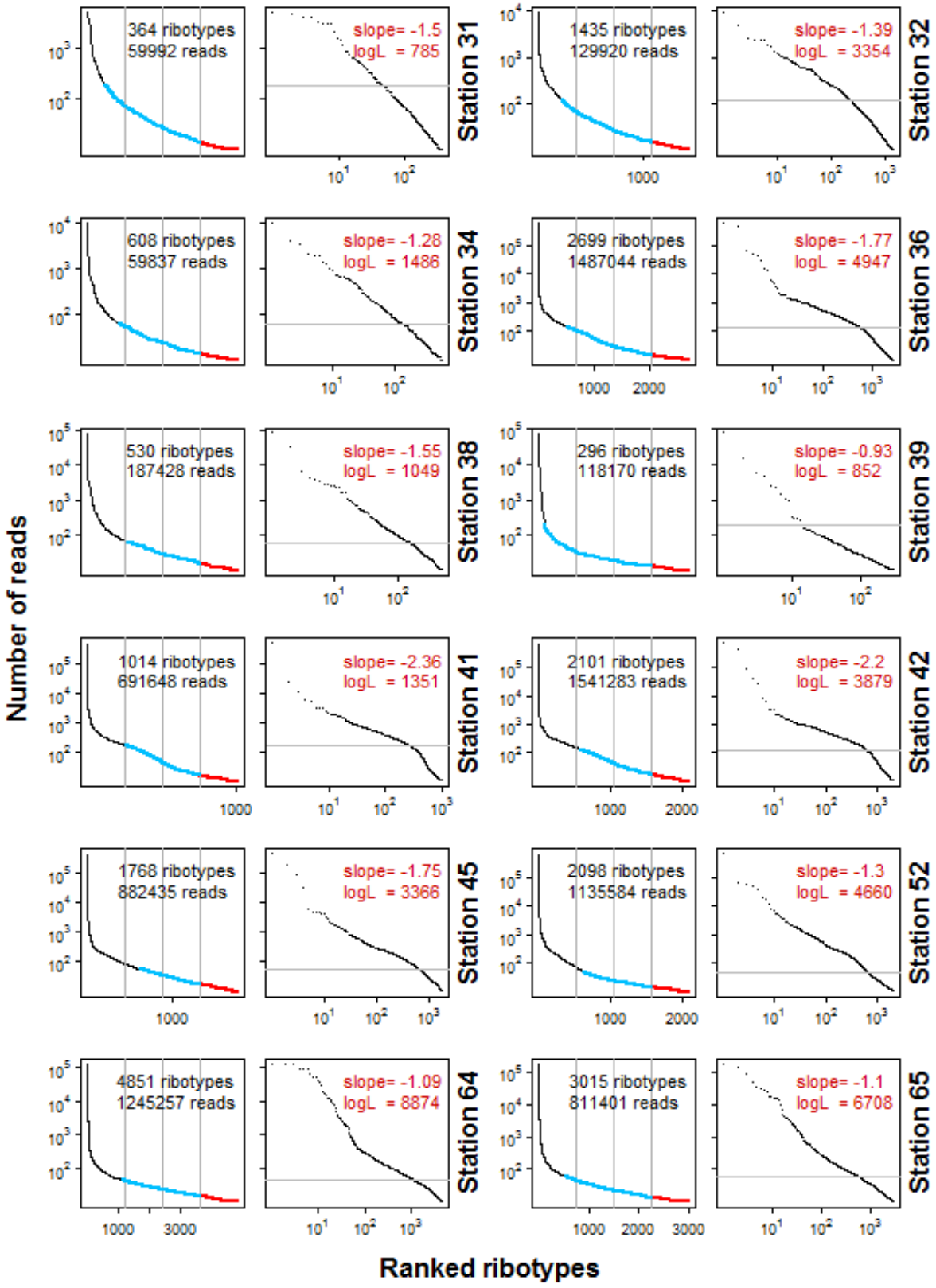
C. Micro-eukaryote (Station 111 to 125)



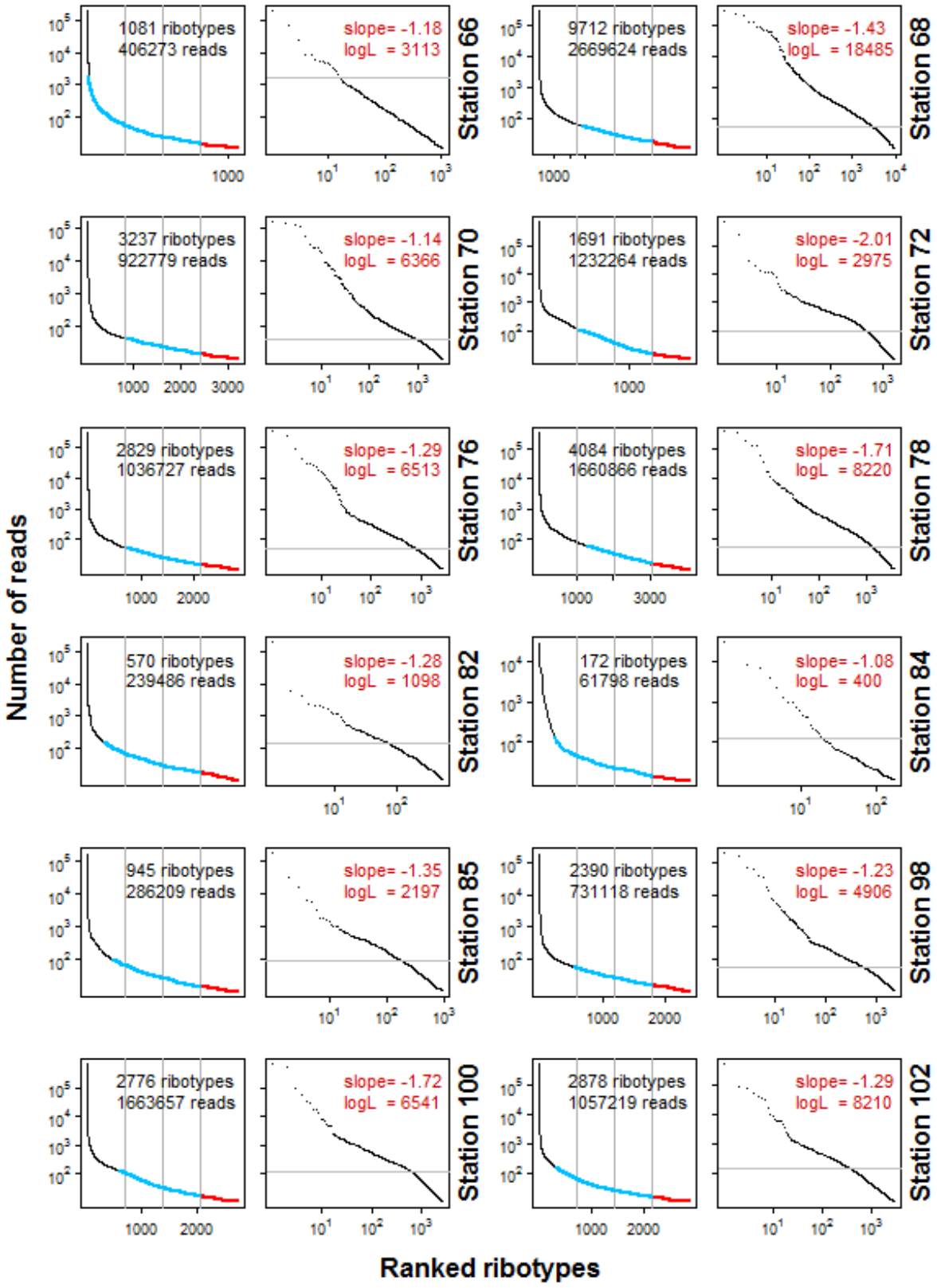
D. Meso-eukaryote (Station 4 to 30)



D. Meso-eukaryote (Station 31 to 65)



D. Meso-eukaryote (Station 66 to 102)



D. Meso-eukaryote (Station 109 to 125)

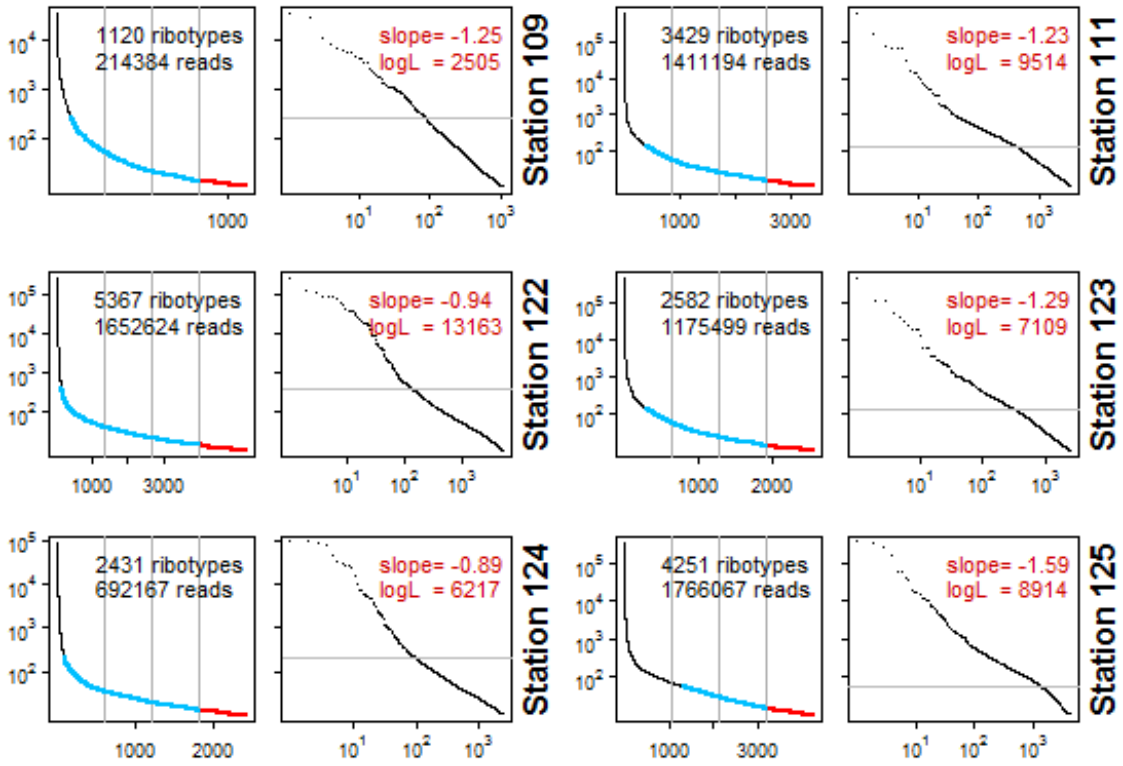


Figure H5.1. Rank-abundance curves of empirical protistan community samples. Abundance in the sample is plotted against ribotype rank in the sample. (A) pico (0.8-5 µm), (B) nano (5-20 µm), (C) micro (20-180 µm) and (D) meso (180-2000 µm).

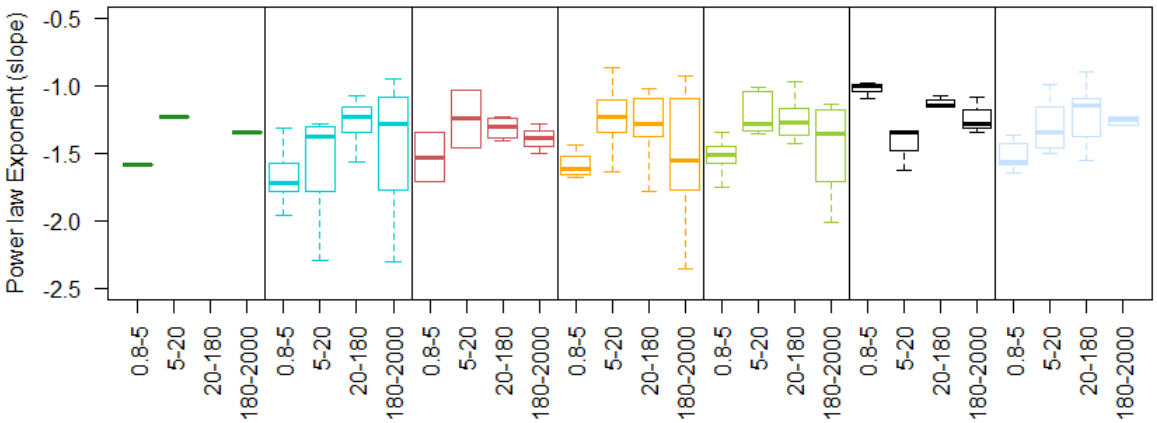


Figure H5.2. Boxplot of slopes clustered based on size (by oceanic provinces).

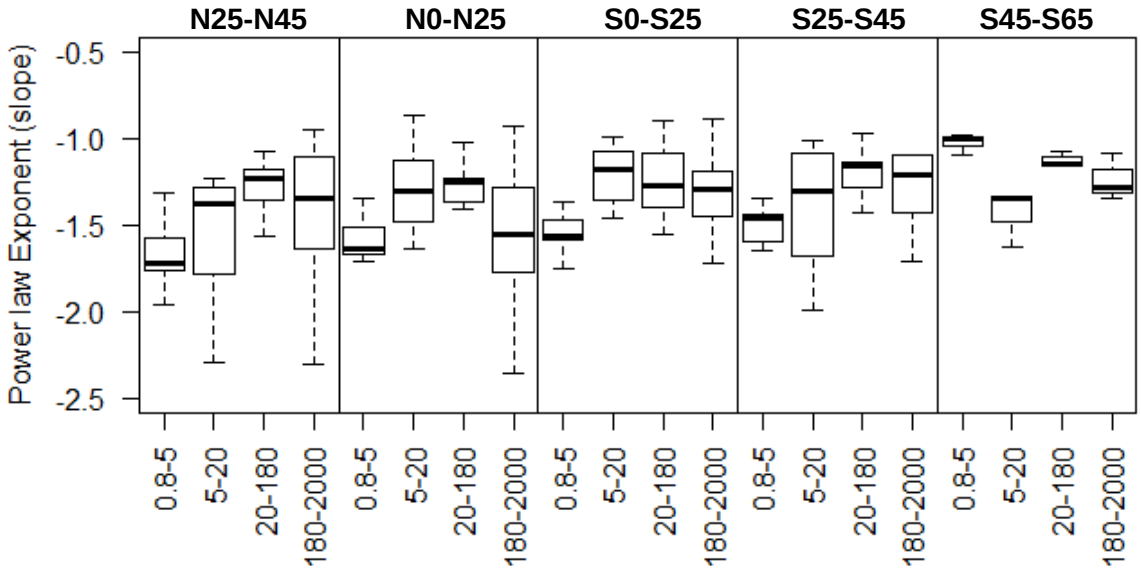


Figure H5.3. Boxplot of slopes clustered based on size (by latitudinal bands).

Table H5.1. Wilcoxon rank sum test statistics comparing the slopes between different size classes combination. P-values for the Mann-Whitney-Wilcoxon Test is reported with the test statistic.

Data 1	Data 2	W	p-value
sur_085	**sur_520**	**569.5**	**0.02686**
sur_085	**sur_20180**	**341.5**	**1.437e-06**
sur_085	**sur_1802000**	**195**	**5.504e-10**
sur_520	sur_20180	679	0.07247
sur_520	**sur_1802000**	**374**	**3.69e-06**
sur_20180	**sur_1802000$**	**460.5**	**1.408e-05**

I. Co-authored manuscripts:

1. *de Vargas et al. (2015) Eukaryotic plankton diversity in the sunlit global ocean. Science: 348(6237) DOI: 10.1126/science.1261605.*

2. *Villar et al. (2015) Environmental characteristics of Agulhas rings affect inter-ocean plankton transport. Science: 348(6237) DOI: 10.1126/science.1261447.*

OCEAN PLANKTON

Eukaryotic plankton diversity in the sunlit ocean

Colomban de Vargas,[1,2]*† Stéphane Audic,[1,2]† Nicolas Henry,[1,2]† Johan Decelle,[1,2]† Frédéric Mahé,[3,1,2]† Ramiro Logares,[4] Enrique Lara,[5] Cédric Berney,[1,2] Noan Le Bescot,[1,2] Ian Probert,[6,7] Margaux Carmichael,[1,2,8] Julie Poulain,[9] Sarah Romac,[1,2] Sébastien Colin,[1,2,8] Jean-Marc Aury,[9] Lucie Bittner,[10,11,8,1,2] Samuel Chaffron,[12,13,14] Micah Dunthorn,[3] Stefan Engelen,[9] Olga Flegontova,[15,16] Lionel Guidi,[17,18] Aleš Horák,[15,16] Olivier Jaillon,[9,19,20] Gipsi Lima-Mendez,[12,13,14] Julius Lukeš,[15,16,21] Shruti Malviya,[8] Raphael Morard,[22,1,2] Matthieu Mulot,[5] Eleonora Scalco,[23] Raffaele Siano,[24] Flora Vincent,[13,8] Adriana Zingone,[23] Céline Dimier,[1,2,8] Marc Picheral,[17,18] Sarah Searson,[17,18] Stefanie Kandels-Lewis,[25,26] *Tara* Oceans Coordinators‡ Silvia G. Acinas,[4] Peer Bork,[25,27] Chris Bowler,[8] Gabriel Gorsky,[17,18] Nigel Grimsley,[28,29] Pascal Hingamp,[30] Daniele Iudicone,[23] Fabrice Not,[1,2] Hiroyuki Ogata,[31] Stephane Pesant,[32,22] Jeroen Raes,[12,13,14] Michael E. Sieracki,[33,34] Sabrina Speich,[35,36] Lars Stemmann,[17,18] Shinichi Sunagawa,[25] Jean Weissenbach,[9,19,20] Patrick Wincker,[9,19,20]* Eric Karsenti[26,8]*

Marine plankton support global biological and geochemical processes. Surveys of their biodiversity have hitherto been geographically restricted and have not accounted for the full range of plankton size. We assessed eukaryotic diversity from 334 size-fractionated photic-zone plankton communities collected across tropical and temperate oceans during the circumglobal *Tara* Oceans expedition. We analyzed 18*S* ribosomal DNA sequences across the intermediate plankton-size spectrum from the smallest unicellular eukaryotes (protists, >0.8 micrometers) to small animals of a few millimeters. Eukaryotic ribosomal diversity saturated at ~150,000 operational taxonomic units, about one-third of which could not be assigned to known eukaryotic groups. Diversity emerged at all taxonomic levels, both within the groups comprising the ~11,200 cataloged morphospecies of eukaryotic plankton and among twice as many other deep-branching lineages of unappreciated importance in plankton ecology studies. Most eukaryotic plankton biodiversity belonged to heterotrophic protistan groups, particularly those known to be parasites or symbiotic hosts.

The sunlit surface layer of the world's oceans functions as a giant biogeochemical membrane between the atmosphere and the ocean interior (*1*). This biome includes plankton communities that fix CO_2 and other elements into biological matter, which then enters the food web. This biological matter can be remineralized or exported to the deeper ocean, where it may be sequestered over ecological to geological time scales. Studies of this biome have typically focused on either conspicuous phyto- or zooplankton at the larger end of the organismal size spectrum or microbes (prokaryotes and viruses) at the smaller end. In this work, we studied the taxonomic and ecological diversity of the intermediate size spectrum (from 0.8 μm to a few millimeters), which includes all unicellular eukaryotes (protists) and ranges from the smallest protistan cells to small animals (*2*). The ecological biodiversity of marine planktonic protists has been analyzed using Sanger (*3*–*5*) and high-throughput (*6*, *7*) sequencing of mainly ribosomal DNA (rDNA) gene markers, on relatively small taxonomic and/or geographical scales, unveiling key new groups of phagotrophs (*8*), parasites (*9*), and phototrophs (*10*). We sequenced 18*S* rDNA metabarcodes up to local and global saturations from size-fractionated plankton communities sampled systematically across the world tropical and temperate sunlit oceans.

A global metabarcoding approach

To explore patterns of photic-zone eukaryotic plankton biodiversity, we generated ~766 million raw rDNA sequence reads from 334 plankton samples collected during the circumglobal *Tara* Oceans expedition (*11*). At each of 47 stations, plankton communities were sampled at two water-column depths corresponding to the main hydrographic structures of the photic zone: subsurface mixed-layer waters and the deep chlorophyll maximum (DCM) at the top of the thermocline. A low-shear, nonintrusive peristaltic pump and plankton nets of various mesh sizes were used on board *Tara* to sample and concentrate appropriate volumes of seawater to theoretically recover complete local eukaryotic biodiversity from four major organismal size fractions: piconanoplankton (0.8 to 5 μm), nanoplankton (5 to 20 μm), microplankton (20 to 180 μm), and mesoplankton (180 to 2000 μm) [see (*12*) for detailed *Tara* Oceans field sampling strategy and protocols].

We extracted total DNA from all samples, polymerase chain reaction (PCR)–amplified the hypervariable V9 region of the nuclear gene that encodes 18*S* rRNA (*13*), and generated an average of 1.73 ± 0.65 million sequence reads (paired-end Illumina) per sample (*11*). Strict bioinformatic quality control led to a final data set of 580 million reads, of which ~2.3 million were distinct,

[1]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [2]Sorbonne Universités, Université Pierre et Marie Curie (UPMC) Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [3]Department of Ecology, University of Kaiserslautern, Erwin-Schroedinger Street, 67663 Kaiserslautern, Germany. [4]Department of Marine Biology and Oceanography, Institute of Marine Science (ICM)–Consejo Superior de Investigaciones Científicas (CSIC), Passeig Marítim de la Barceloneta 37-49, Barcelona E08003, Spain. [5]Laboratory of Soil Biology, University of Neuchâtel, Rue Emile-Argand 11, 2000 Neuchâtel, Switzerland. [6]CNRS, FR2424, Roscoff Culture Collection, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [7]Sorbonne Universités, UPMC Paris 06, FR 2424, Roscoff Culture Collection, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [8]Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), and Inserm U1024, and CNRS UMR 8197, Paris, F-75005 France. [9]Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91000 Evry, France. [10]CNRS FR3631, Institut de Biologie Paris-Seine, F-75005, Paris, France. [11]Sorbonne Universités, UPMC Paris 06, Institut de Biologie Paris-Seine, F-75005, Paris, France. [12]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [13]Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. [14]Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [15]Institute of Parasitology, Biology Centre, Czech Academy of Sciences, Branišovská 31, 37005 České Budějovice, Czech Republic. [16]Faculty of Science, University of South Bohemia, Branišovská 31, 37005 České Budějovice, Czech Republic. [17]CNRS, UMR 7093, Laboratoire d'Océanographie de Villefranche-sur-Mer (LOV), Observatoire Océanologique, F-06230, Villefranche-sur-Mer, France. [18]Sorbonne Universités, UPMC Paris 06, UMR 7093, LOV, Observatoire Océanologique, F-06230, Villefranche-sur-Mer, France. [19]CNRS, UMR 8030, CP5706, Evry, France. [20]Université d'Evry, UMR 8030, CP5706, Evry, France. [21]Canadian Institute for Advanced Research, 180 Dundas Street West, Suite 1400, Toronto, Ontario M5G 1Z8, Canada. [22]MARUM, Center for Marine Environmental Sciences, University of Bremen, 28359 Bremen, Germany. [23]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [24]Ifremer, Centre de Brest, DYNECO/Pelagos CS 10070, 29280 Plouzané, France. [25]Structural and Computational Biology, European Molecular Biology Laboratory (EMBL), Meyerhofstraße 1, 69117 Heidelberg, Germany. [26]Directors' Research, EMBL, Meyerhofstraße 1, 69117 Heidelberg, Germany. [27]Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany. [28]CNRS UMR 7232, Biologie Intégrative des Organismes Marins (BIOM), Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [29]Sorbonne Universités Paris 06, Observatoire Océanologique de Banyuls (OOB) UPMC, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [30]Aix Marseille Université, CNRS IGS UMR 7256, 13288 Marseille, France. [31]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan. [32]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany. [33]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME 04544, USA. [34]National Science Foundation, Arlington, VA 22230, USA. [35]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France. [36]Laboratoire de Physique des Océans, Université de Bretagne Occidentale (UBO)–Institut Universitaire Européen de la Mer (IUEM), Place Copernic, 29820 Plouzané, France.

***Corresponding author. E-mail: vargas@sb-roscoff.fr (C.d.V.); pwincker@genoscope.cns.fr (P.W.); karsenti@embl.de (E.K.)**

†These authors contributed equally to this work. ‡*Tara* Oceans Coordinators and affiliations appear at the end of this paper.

hereafter denoted "metabarcodes." We then clustered metabarcodes into biologically meaningful operational taxonomic units (OTUs) (*14*) and assigned a eukaryotic taxonomic path to all metabarcodes and OTUs by global similarity analysis with 77,449 reference, Sanger-sequenced V9 rDNA barcodes covering the known diversity of eukaryotes and assembled into an in-house database called *V9_PR2* (*15*). Beyond taxonomic assignation, we inferred basic trophic and symbiotic ecological modes (photo- versus heterotrophy; parasitism, commensalism, mutualism for both hosts and symbionts) to *Tara* Oceans reads and OTUs on the basis of their genetic affiliation to large monophyletic and monofunctional groups of reference barcodes. We finally inferred large-scale ecological patterns of eukaryotic biodiversity across geography, taxonomy, and organismal size fractions based on rDNA abundance data and community similarity analyses and compared them to current knowledge extracted from the literature.

The extent of eukaryotic plankton diversity in the photic zone of the world ocean

Sequencing of ~1.7 million V9 rDNA reads from each of the 334 size-fractionated plankton samples was sufficient to approach saturation of eukaryotic richness at both local and global scales (Fig. 1, A and B). Local richness represented, on average, 9.7 ± 4% of global richness, the latter approaching saturation at ~2 million eukaryotic metabarcodes or ~110,000 OTUs (*16*). The global pool of OTUs displayed a good fit to the truncated Preston log-normal distribution (*17*), which, by extrapolation, suggests a total photic-zone eukaryotic plankton richness of ~150,000 OTUs, of which ~40,000 were not found in our survey (Fig. 1C). Thus, we estimate that our survey unveiled ~75% of eukaryotic ribosomal diversity in the globally distributed water masses analyzed. The extrapolated ~150,000 total OTUs is much higher than the ~11,200 formally described species of marine eukaryotic plankton (see below) and probably represents a highly conservative, lower-boundary estimate of the true number of eukaryotic species in this biome, given the relatively limited taxonomic resolution power of the 18*S* rDNA gene. Our data indicate that eukaryotic taxonomic diversity is higher in smaller organismal size fractions, with a peak in the piconanoplankton (Fig. 1A), highlighting the richness of tiny organisms that are poorly characterized in terms of morphotaxonomy and physiology (*18*). A first-order, supergroup-level classification of all *Tara* Oceans OTUs demonstrated the prevalence (at the biome scale and across the >four orders of size magnitude sampled) of protist rDNA biodiversity with respect to that of classical multicellular eukaryotes, i.e., animals, plants, and fungi (Fig. 2A). Protists accounted for >85% of total eukaryotic ribosomal diversity, a ratio that may well hold true for other marine, freshwater, and terrestrial oxygenic ecosystems (*19*). The latest estimates of total marine eukaryotic biodiversity based on statistical extrapolations from classical taxonomic knowledge predict the existence of 0.5 to 2.2 million species [including all benthic and planktonic systems from reefs to deep-sea vents (*20*, *21*)] but do not take into account the protistan knowledge gap highlighted here. Simple application of our animal–to–other eukaryotes ratio of ~13% to the robust prediction of the total number of metazoan species from (*20*) would imply that 16.5 million and 60 million eukaryotic species potentially inhabit the oceans and Earth, respectively.

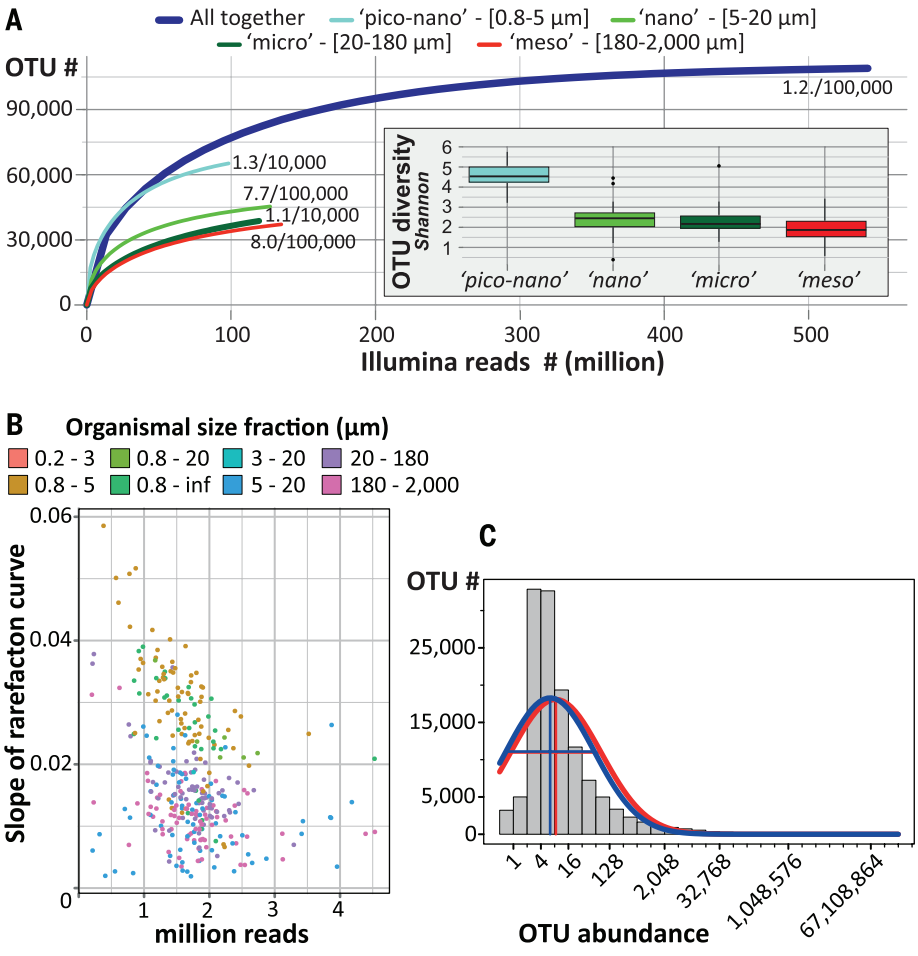


Fig. 1. Photic-zone eukaryotic plankton ribosomal diversity. (**A**) V9 rDNA OTUs rarefaction curves and overall diversity (Shannon index, inset) for each plankton organismal size fraction. Proximity to saturation is indicated by weak slopes at the end of each rarefaction curve (e.g., 1.2/100,000 means 1.2 novel metabarcodes obtained every 100,000 rDNA reads sequenced). (**B**) Saturation slope versus number of V9 rDNA reads for all of the 334 samples (dots) analyzed herein. A slope of 0.02 indicates that two novel barcodes can be recovered if 100 new reads are sequenced. Samples are colored according to size fraction. (**C**) Global OTU abundance distribution and fit to the Preston log-normal model. Most OTUs in our data set were represented by 3 to 16 reads, whereas fewer OTUs presented less or more abundances. Quasi-Poisson fit to octaves (red curve) and maximized likelihood to $\log_2$ abundances (blue curve) approximations were used to fit the OTU abundance distribution to the Preston log-normal model. Overall, the global (A) and local (B) saturation values indicate that our extensive sampling effort (in terms of spatiotemporal coverage and sequencing depth) uncovered the majority of eukaryotic ribosomal diversity within the photic layer of the world's tropical to temperate oceans. Calculation of the Preston veil, which infers the number of OTUs that we missed (or were veiled) during our sampling (~40,000), confirmed that we captured most of the protistan richness, thus allowing extraction of holistic and general patterns of eukaryotic plankton biodiversity from our data set.

Phylogenetic breakdown of photic-zone eukaryotic biodiversity

About one-third of eukaryotic ribosomal diversity in our data set did not match any reference barcode in the extensive *V9_PR2* database ("unassigned" category in Fig. 2A). This unassignable diversity represented only a small proportion (2.6%) of total reads and increased in both richness and abundance in smaller organismal size fractions, suggesting that it corresponds mostly to rare and minute taxa that have escaped previous characterization. Some may also correspond to divergent rDNA pseudogenes, known to exist in eukaryotes (*22*, *23*) or sequencing artefacts (*24*), although both of these would be expected to be present in equal proportion in all

size fractions [details in (*16*)]. The remaining ~87,000 assignable OTUs were classified into 97 deep-branching lineages covering the full spectrum of cataloged eukaryotic diversity amongst the seven recognized supergroups and multiple lineages of uncertain placement (*15*) whose origins go back to the primary radiation of eukaryotic life in the Neoproterozoic. Although highly represented in the *V9_PR2* reference database, several well-known lineages adapted to terrestrial, marine benthic, or anaerobic habitats (e.g., Embryophyta; apicomplexan and trypanosome parasites of land plants and animals; amoeboflagellate Breviatea; and several lineages of Amoebozoa, Excavata, and Cercozoa) were not detected in our metabarcoding data set, suggesting the absence of contamination during the PCR and sequencing steps on land and reducing the number of deep branches of eukaryotic plankton to 85 (Fig. 3).

We then extracted the metabarcodes assigned to morphologically well-known planktonic eukaryotic taxa from our data set and compared them with the conventional, 150 year-old morphological view of marine eukaryotic plankton that includes ~11,200 cataloged species divided into three broad categories: ~4350 species of phytoplankton (microalgae), ~1350 species of protozooplankton (relatively large, often biomineralized, heterotrophic protists), and ~5500 species of metazooplankton (holoplanktonic animals) (*25*–*27*). A congruent picture of the distribution of morphogenetic diversity among and within these organismal categories emerged from our data set (Fig. 2B), but typically, three to eight times more rDNA OTUs were found than described morphospecies in the best-known lineages within these categories. This is within the range of the number of cryptic species typically detected in globally-distributed pelagic taxa using molecular data (*28*, *29*). The general congruency between genetic and morphological data in the cataloged compartment of eukaryotic plankton suggests that the protocols used, from plankton sampling to DNA sequencing, recovered the known eukaryotic biodiversity without major qualitative or quantitative biases. However, OTUs related to morphologically described taxa represented only a minor part of the total eukaryotic plankton ribosomal and phylogenetic diversity. Overall, <1% of OTUs were strictly identical to reference sequences, and OTUs were, on average, only ~86% similar to any V9 reference sequence (Fig. 3F) (*16*). This shows that most photic-zone eukaryotic plankton V9 rDNA diversity had not been previously sequenced from cultured strains, single-cell isolates, or even environmental clone library surveys. The *Tara* Oceans metabarcode data set added considerable phylogenetic information to previous protistan rDNA knowledge, with an estimated mean tree-length increase of 453%, reaching >100% in 43 lineages (*16*). Even in the best-referenced groups such as the diatoms (1232 reference sequences) (Fig. 3B), we identified many new rDNA sequences, both within known groups and forming new clades (*16*).

Eleven "hyperdiverse" lineages each contained >1000 OTUs, together representing ~88 and

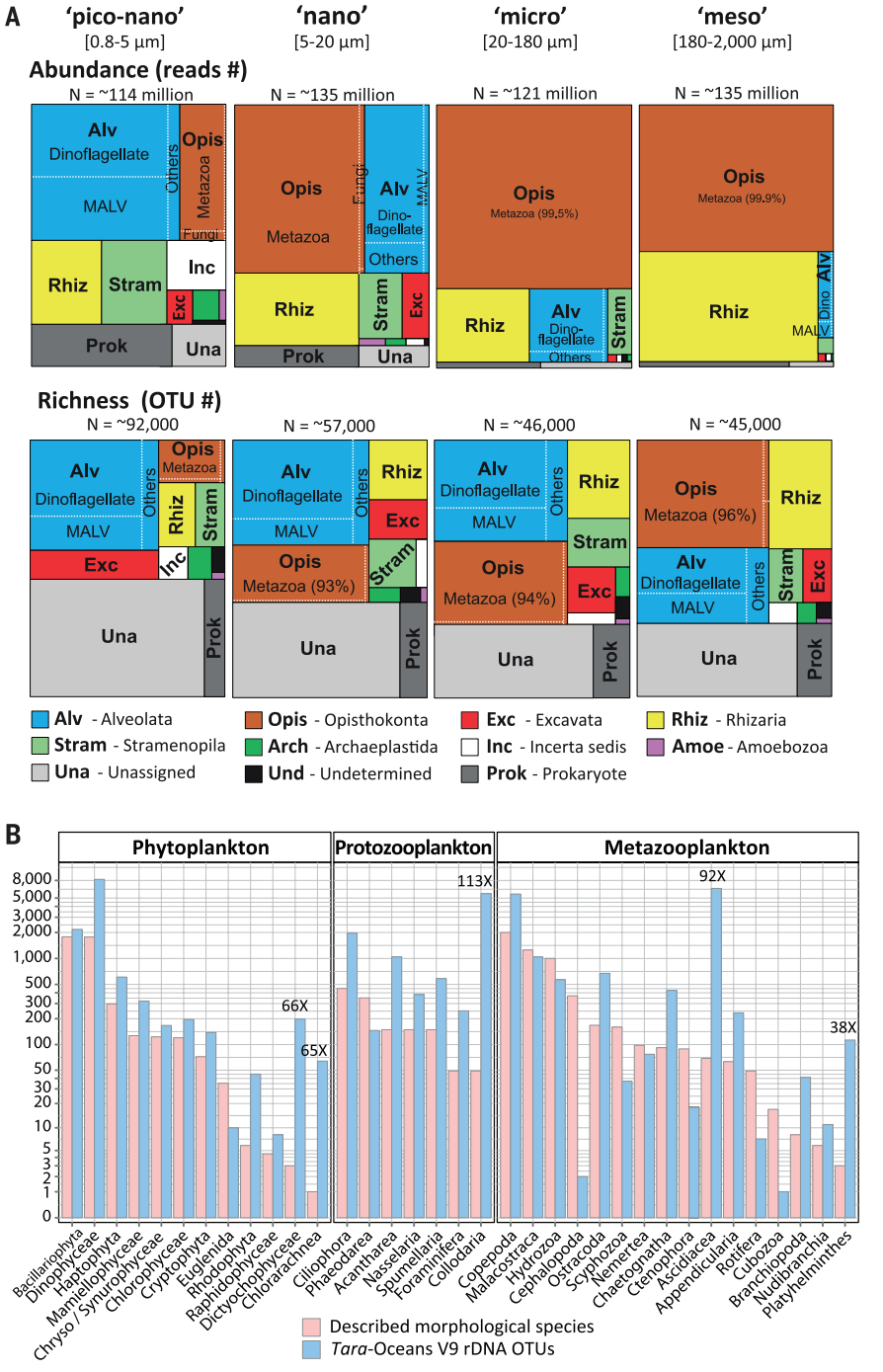


Fig. 2. Unknown and known components of eukaryotic plankton biodiversity. (**A**) Phylogenetic breakdown of the entire metabarcoding data set at the eukaryotic supergroup level. All *Tara* Oceans V9 rDNA reads and OTUs were classified among the seven recognized eukaryotic supergroups plus the known but unclassified deep-branching lineages (incertae sedis). The tree maps display the relative abundance (upper part) and richness (lower part) of the different eukaryotic supergroups in each organismal size fraction. Note that ~5% of barcodes were assigned to prokaryotes, essentially in the piconano fraction, witnessing the universality of the eukaryotic primers used. Barcodes are "unassigned" when sequence similarity to a reference sequence is <80% and "undetermined" when eukaryotic supergroups could not be discriminated (at similarity >80%). (**B**) Ribosomal DNA diversity associated with the morphologically known and cataloged part of eukaryotic plankton. The total number of morphologically described species in the literature [red bars, based on (*25*–*27*)] and the corresponding total number of *Tara* Oceans V9 rDNA OTUs (blue bars) are indicated for each of the 35 classical lineages of eukaryotic phyto-, protozoo-, and metazooplankton. The five classical groups that were found to be substantially more diverse than previously thought (from 38- to 113-fold more OTUs than morphospecies) are highlighted. Note that in the classical morphological view, phyto- and metazooplankton comprise ~88% of total eukaryotic plankton diversity.

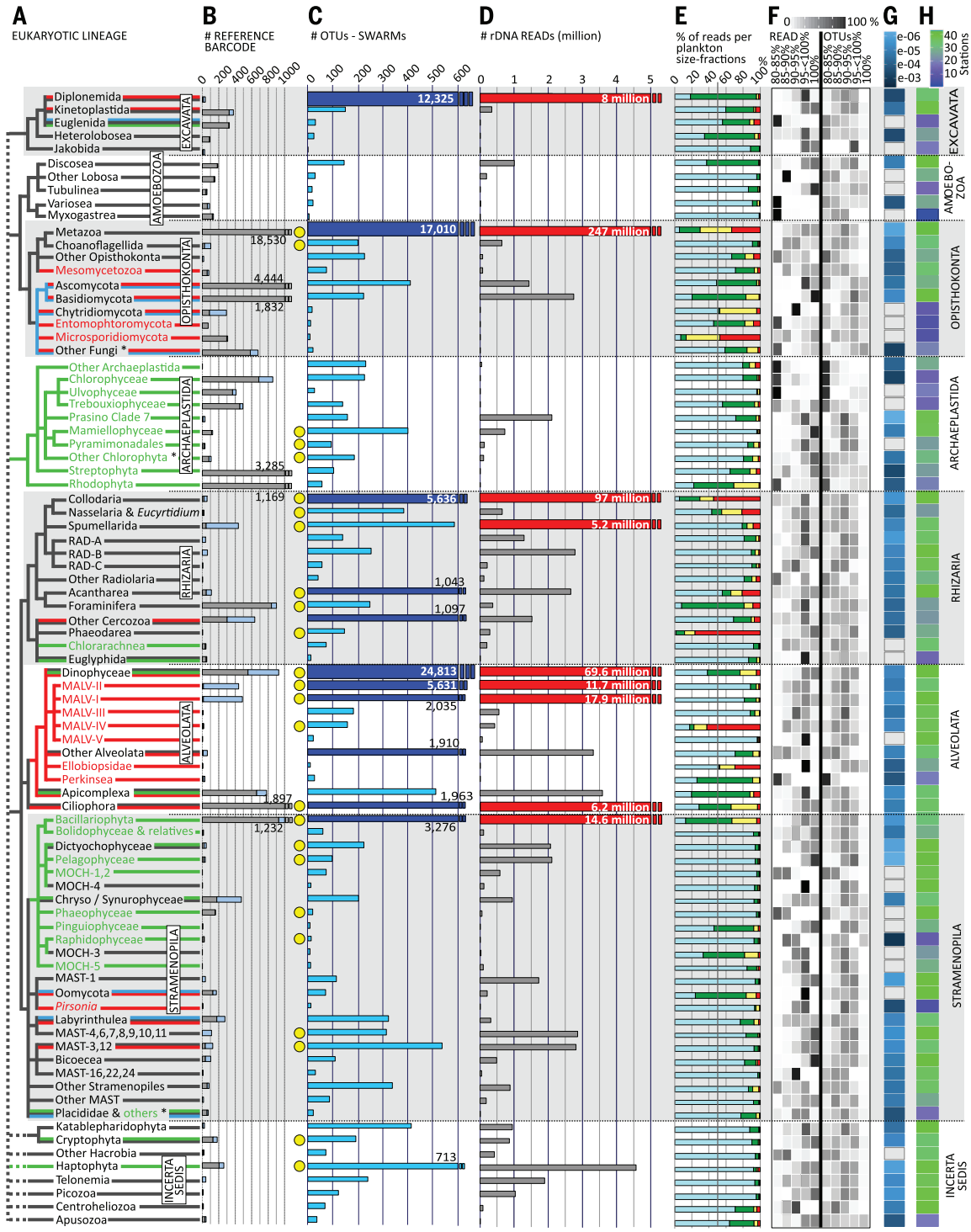


Fig. 3. Phylogenetic distribution of the assignable component of eukaryotic plankton ribosomal diversity. (**A**) Schematic phylogeny of the 85 deep-branching eukaryotic lineages represented in our global oceans metabarcoding data set, with broad ecological traits based on current knowledge: red, parasitic; green, photoautotrophic; blue, osmo- or saprotrophic; black, mostly phagotrophic lineages. Lineages known only from environmental sequence data were colored in black by default. For simplicity, three branches (denoted by asterisks) artificially group a few distinct lineages [details in (*15*)]. (**B**) Number of reference V9 rDNA barcodes used to annotate the metabarcoding data set (gray, with known taxonomy at the genus and/or species level; light blue, from previous 18S rDNA environmental clone libraries). (**C**) *Tara* Oceans V9 rDNA OTU richness. Dark blue thicker bars indicate the 11 hyperdiverse lineages containing >1000 OTUs. Yellow circles highlight the 25 lineages that have been recognized as important in previous marine plankton biodiversity and ecology studies using morphological and/or molecular data [see also (*15*)]. (**D**) Eukaryotic plankton abundance expressed as numbers of rDNA reads (the red bars indicate the nine most abundant lineages with >5 million reads). (**E**) Proportion of rDNA reads per organismal size fraction. Light blue, piconano-; green, nano-; yellow, micro-; red, mesoplankton. (**F**) Percentage of reads and OTUs with 80 to 85%, 85 to 90%, 90 to 95%, 95 to <100%, and 100% sequence similarity to a reference sequence. (**G**) Slope of OTU rarefaction curves. (**H**) Mean geographic occupancy (average number of stations in which OTUs were observed, weighted by OTU abundance).

~90% of all OTUs and reads, respectively (Fig. 3C). Among these, the only permanently phototrophic taxa were diatoms (Fig. 4A) and about one-third of dinoflagellates (Fig. 4, B to F), together comprising ~15 and ~13% of hyperdiverse OTUs and reads, respectively (*30*). Most hyperdiverse photic-zone plankton belonged to three supergroups—the Alveolata, Rhizaria, and Excavata—about which we have limited biological or ecological information. The Alveolata, which consist mostly of parasitic [marine alveolates (MALVs)] (Fig. 4F) and phagotrophic (ciliates and most dinoflagellates) taxa, were by far the most diverse supergroup, comprising ~42% of all assignable OTUs. The Rhizaria are a group of amoeboid heterotrophic protists with active pseudopods displaying a broad spectrum of ecological behavior, from phagotrophy to parasitism and mutualism (symbioses) (*31*). Rhizarian diversity peaked in the Retaria (Fig. 4, C and D) a subgroup including giant protists that build complex skeletons of silicate (Polycystinea), strontium sulfate (Acantharia) (Fig. 4C), or calcium carbonate (Foraminifera) and thus comprise key microfossils for paleoceanography. Unsuspected rDNA diversity was recorded within the Collodaria (5636 OTUs), polycystines that are mostly colonial, poorly silicified, or naked and live in obligatory symbiosis with photosynthetic dinoflagellates (Fig. 4D) (*32*, *33*). Arguably, the most surprising component of novel biodiversity was the >12,300 OTUs related to reference sequences of diplonemids, an excavate lineage that has only two described genera of flagellate grazers, one of which parasitizes diatoms and crustaceans (*34*, *35*). Their ribosomal diversity was not only much higher than that observed in classical plankton groups such as foraminifers, ciliates, or diatoms (50-fold, 6-fold, and 3.8-fold higher, respectively) but was also far from richness saturation (Fig. 3E). Eukaryotic rDNA diversity peaked especially in the few lineages that extend across larger size fractions (i.e., metazoans, rhizarians, dinoflagellates, ciliates, diatoms) (Fig. 3E). Larger cells or colonies not only provide protection against predation via size-mediated avoidance and/or construction of composite skeletons but also provide support for complex and coevolving relationships with often specialized parasites or mutualistic symbionts.

Beyond this hyperdiverse, largely heterotrophic eukaryotic majority, our data set also highlighted the phylogenetic diversity of poorly known phagotrophic (e.g., 413 OTUs of Katablepharidophyta, 240 OTUs of Telonemia), osmotrophic (e.g., 410 OTUs of Ascomycota, 322 OTUs of Labyrinthulea), and parasitic (e.g., 384 OTUs of gregarine apicomplexans, 160 OTUs of Ascetosporea, 68

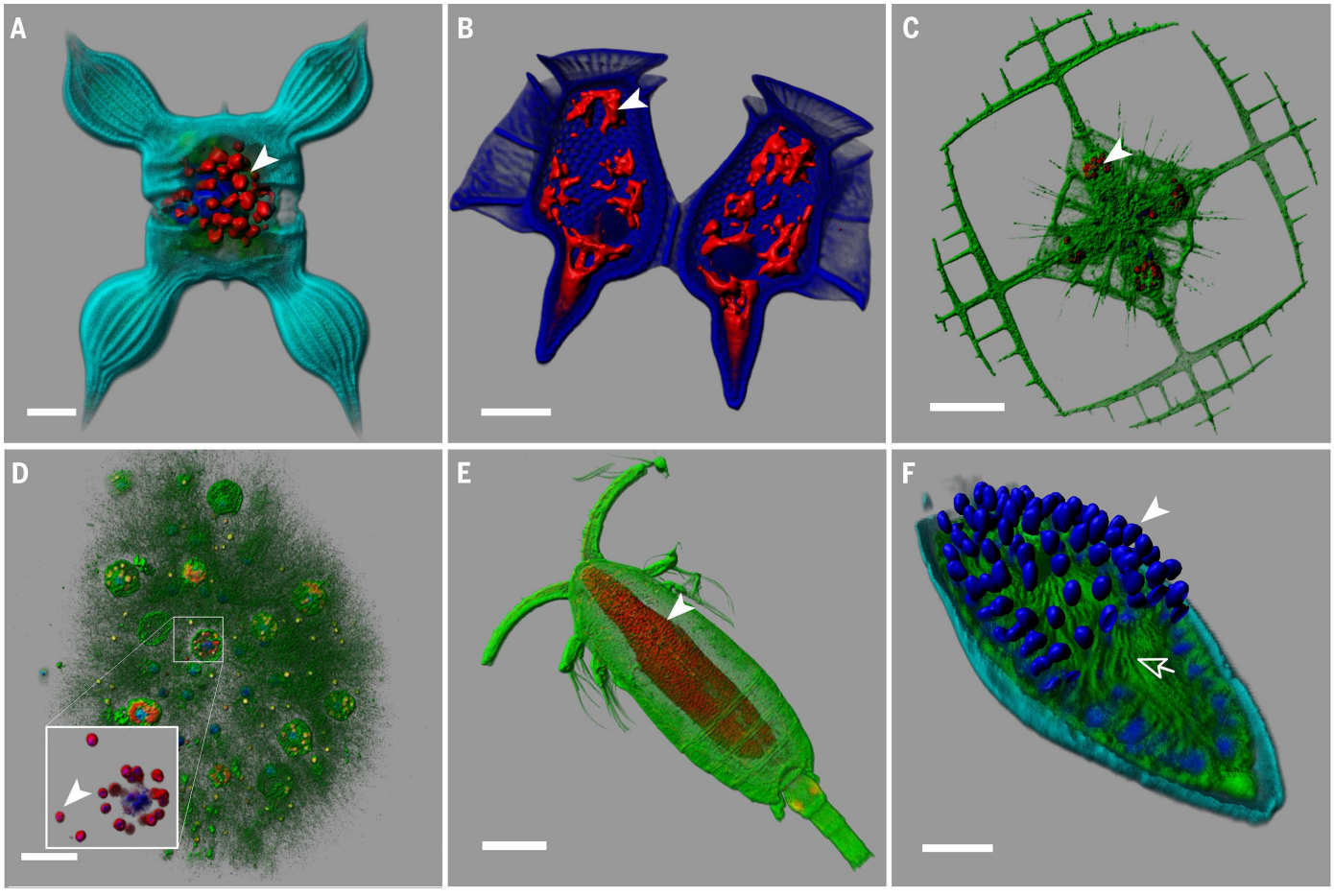


Fig. 4. Illustration of key eukaryotic plankton lineages. (**A**) Stramenopila; a phototrophic diatom *Chaetoceros bulbosus*, with its chloroplasts in red (arrowhead). Scale bar, 10 μm. (**B**) Alveolata; a heterotrophic dinoflagellate *Dinophysis caudata* harboring kleptoplasts [in red (arrowhead)]. Scale bar, 20 μm (*75*). (**C**) Rhizaria; an acantharian *Lithoptera* sp. with endosymbiotic haptophyte cells from the genus *Phaeocystis* [in red (arrowhead)]. Scale bar, 50 μm (*41*). (**D**) Rhizaria; inside a colonial network of Collodaria, a cell surrounded by several captive dinoflagellate symbionts of the genus *Brandtodinium* (arrowhead). Scale bar, 50 μm (*33*). (**E**) Opisthokonta; a copepod whose gut is colonized by the parasitic dinoflagellate *Blastodinium* [red area shows nuclei (arrowhead)]. Scale bar, 100 μm (*51*). (**F**) Alveolata; a cross-sectioned, dinoflagellate cell infected by the parasitoid alveolate *Amoebophrya* (MALV-II). Each blue spot (arrowhead) is the nucleus of future free-living dinospores; their flagella are visible in green inside the mastigocoel cavity (arrow). Scale bar, 5 μm. The cellular membranes were stained with DiOC6 (green); DNA and nuclei were stained with Hoechst (blue) [the dinoflagellate theca in (B) was also stained by this dye]. Chlorophyll autofluorescence is shown in red [except for in (E)]. An unspecific fluorescent painting of the cell surface (light blue) was used to reveal cell shape for (A) and (F). All specimens come from *Tara* Oceans samples preserved for confocal laser scanning fluorescent microscopy. Images were three-dimensionally reconstructed with Imaris (Bitplane).

OTUs of Ichthyosporea) protist groups. Amongst the 85 major lineages presented in the phylogenetic framework of Fig. 3, less than one-third (~25) have been recognized as important in previous marine plankton biodiversity and ecology studies using morphological and/or molecular data (Fig. 3C) (*15*). The remaining ~60 branches had either never been observed in marine plankton or were detected through morphological description of one or a few species and/or the presence of environmental sequences in geographically restricted clone library surveys (*15*). This understudied diversity represents ~25% of all taxonomically assignable OTUs (>21,500) and covers broad taxonomic and geographic scales, thus representing a wealth of new actors to integrate into future plankton systems biology studies.

Insights into photic-zone eukaryotic plankton ecology

Functional annotation of taxonomically assigned V9 rDNA metabarcodes was used as a first attempt to explore ecological patterns of eukaryotic diversity across broad spatial scales and organismal size fractions, focusing on fundamental trophic modes (photo- versus heterotrophy) and symbiotic interactions (parasitism to mutualism). Heterotroph (protists and metazoans) V9 rDNA metabarcodes were substantially more diverse (63%) and abundant (62%) than phototroph metabarcodes that represented <20% of OTUs and reads across all size fractions and geographic sites, with an increasing heterotroph-to-phototroph ratio in the micro- and mesoplankton (Fig. 5A, confirmed in 17 non–size-fractionated samples (*30*). These results challenge the classical morphological view of plankton diversity, biased by a terrestrial ecology approach, whereby phyto- and metazooplankton (the plant-animal paradigm) are thought to comprise ~88% of eukaryotic plankton diversity (Fig. 2B) and heterotrophic protists are typically reduced in food-web modeling to a single entity, often idealized as ciliate grazers.

An unsuspected richness and abundance of metabarcodes assigned to monophyletic groups of heterotrophic protists that cannot survive without endosymbiotic microalgae was found in larger size fractions ("photosymbiotic hosts" in Fig. 5A). Their abundance and even diversity were sometimes greater than those of all metazoan metabarcodes, including those from copepods. Most of these cosmopolitan photosymbiotic hosts were found within the hyperdiverse radiolarians Acantharia (1043 OTUs) and Collodaria (5636 OTUs) (Figs. 3, 4B, and 5D), which have often been overlooked in traditional morphological surveys of plankton-net–collected material because of their delicate gelatinous and/or easily dissolved structures but are known to be very abundant from microscope-based and in situ imaging studies (*36*–*38*). All 95 known colonial collodarian species described since the 19th century (*39*) harbor intracellular symbiotic microalgae, and these key players for plankton ecology are protistan analogs of photosymbiotic corals in tropical coastal reef ecosystems with no equivalent in terrestrial ecology. In addition to their contribution to total primary production (*36*, *38*), these diverse, biologically complex, often biomineralized, and relatively long-lived giant mixotrophic protists stabilize carbon in larger size fractions and probably increase its flux to the ocean interior (*38*). Conversely, the microalgae that are known obligate intracellular partners in open-ocean photosymbioses (*33*, *40*–*42*) (Fig. 5B) were neither very diverse nor highly abundant and occurred evenly across organismal size fractions (Fig. 5C). However, their relative contribution was greatest in the mesoplankton category (10%) (Fig. 5C), where the known photosymbionts of pelagic rhizarians were found (together with their hosts) (Fig. 5B). The stable and systematic abundance of photosymbiotic microalgae across size fractions [a pattern not shown by nonphotosymbiotic microalgae (*30*)] suggests that pelagic photosymbionts maintain free-living and potentially actively growing populations in the piconano- and nanoplankton, representing an accessible pool for recruitment by their heterotrophic hosts. This appears to contrast with photosymbioses in coral reefs and terrestrial systems, where symbiotic microalgal populations mainly occur within their multicellular hosts (*43*).

On the other end of the spectrum of biological interactions, rDNA metabarcodes affiliated to groups of known parasites were ~90 times more diverse than photosymbionts in the piconanoplankton, where they represented ~59% of total heterotrophic protistan ribosomal richness and ~53% of abundance (Figs. 4 and 5C), although this latter value may be inflated by a hypothetically higher rDNA copy number in some marine alveolate lineages (*18*). Parasites in this size fraction were mostly (89% of diversity and 88% of abundance across all stations) within the MALV-I and -II Syndiniales (*30*), which are known exclusively as parasitoid species that kill their hosts and release hundreds of small (2 to 10 μm), nonphagotrophic dinospores (*9*, *44*) that survive for only a few days in the water column (*45*). Abundant parasite-assigned metabarcodes in small size fractions (Fig. 5, B and C) suggest the existence of a large and diverse pool of free-living parasites in photic-zone piconanoplankton, mirroring phage ecology (*46*) and reflecting the extreme diversity and abundance of their known main hosts: radiolarians, ciliates, and dinoflagellates (Fig. 3) (*9*, *47*–*49*). Contrasting with the pattern observed for metabarcodes affiliated to purely phagotrophic taxa, the relative abundance and richness of putative parasite metabarcodes decreased in the nano- and microplanktonic size fractions but increased again in the mesoplankton (Fig. 5C), where parasites are most likely in their infectious stage within larger-sized host organisms. This putative in hospite parasites richness, equivalent to only 23% of that in the piconanoplankton, consisted mostly of a variety of alveolate taxa known to infect crustaceans: MALV-IV such as *Haematodinium* and *Syndinium*; dinoflagellates such as *Blastodinium* (Fig. 4E); and apicomplexan gregarines, mainly *Cephaloidophoroidea* (Fig. 5B) (*9*, *50*, *51*). This pattern contrasts with terrestrial systems where most parasites live within their hosts and are typically transmitted either vertically or through vectors because they generally do not survive outside their hosts (*52*). In the pelagic realm, free-living parasitic spores, like phages, are protected from dessication and dispersed by water diffusion and are apparently massively produced, which likely increases horizontal transmission rate.

Community structuring of photic-zone eukaryotic plankton

Clustering of communities by their compositional similarity revealed the primary influence of organism size ($P = 10^{-3}$, $r^2 = 0.73$) on community structuring, with piconanoplankton displaying stronger cohesiveness than larger organismal size fractions (Fig. 6A). Filtered size-fraction–specific communities separated by thousands of kilometers were more similar in composition than they were to communities from other size fractions at the same location. This was emphasized by the fact that ~36% of all OTUs were restricted to a single size category (*53*). Further analyses within each organismal size fraction indicated that geography plays a role in community structuring, with samples being partially structured according to basin of origin, a pattern that was stronger in larger organismal size fractions ($P = 0.001$ in all cases, $r^2 = 0.255$ for piconanoplankton, 0.371 for nanoplankton, 0.473 for microplankton, and 0.570 for mesoplankton) (Fig. 6B). Mantel correlograms comparing Bray-Curtis community similarity to geographic distances between all samples indicated significant positive correlations in all organismal size fractions over the first ~6000 km, the correlation breaking down at larger geographic distances (*54*). This positive correlation between community dissimilarity and geographic distance, expected under neutral biodiversity dynamics (*55*), challenges the classical niche model for photic-zone eukaryotic plankton biogeography (*56*). The significantly stronger community differentiation by ocean basin in larger organismal size fractions (Fig. 6B) suggests increasing dispersal limitation from piconano- to nano-, micro-, and mesoplankton. Thus, larger-sized eukaryotic plankton communities, containing the highest abundance and diversity of metazoans (Figs. 2A and 5B), were spatially more heterogeneous in terms of both taxonomic (Fig. 6) and functional (Fig. 5A) composition and abundance. The complex life cycle and behaviors of metazooplankton, including temporal reproductive and growth cycles and vertical migrations, together with putative rapid adaptive evolution processes to mesoscale oceanographic features (*57*), may explain the stronger geographic differentiation of mesoplanktonic communities. By contrast, eukaryotic communities in the piconanoplankton were richer (Fig. 1A) and more homogeneous in taxonomic composition (Fig. 6), representing a stable compartment across the world's oceans (*58*).

Even though protistan communities were diverse, the proportions of abundant (>1%) and

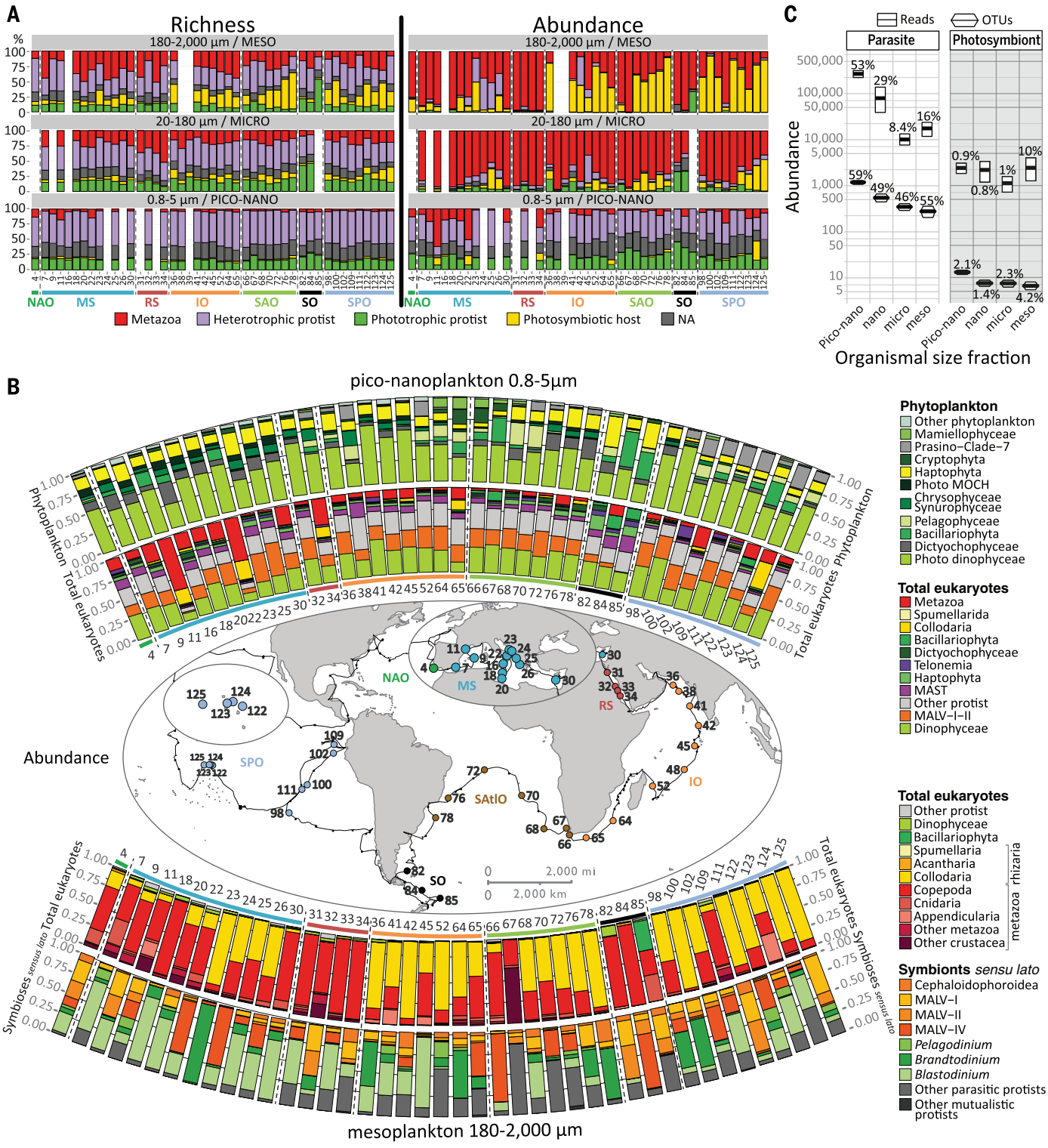


Fig. 5. Metabarcoding inference of trophic and symbiotic ecological diversity of photic-zone eukaryotic plankton. (**A**) Richness (OTU number) and abundance (read number) of rDNA metabarcodes assigned to various trophic taxo-groups across plankton organismal size fractions and stations. Note that the nano size fraction did not contain enough data to be used in this biogeographical analysis [for all size-fraction data, see (*30*)]. NA, not applicable. (**B**) Relative abundance of major eukaryotic taxa across *Tara* Oceans stations for (i) phytoplankton and all eukaryotes in piconanoplankton (above the map) and (ii) all eukaryotes and protistan symbionts (sensu lato) in mesoplankton (below the map). Note the pattern of inverted relative abundance between collodarian colonies (Fig. 4) and copepods in, respectively, the oligotrophic and eutrophic and mesotrophic systems. The dinoflagellates *Brandtodinium* and *Pelagodinium* are endophotosymbionts in Collodaria (*33*) and Foraminifera (*40*, *42*), respectively. (**C**) Richness and abundance of parasitic and photosymbiotic (microalgae) protists across organismal size fractions. The relative contributions (percent) of parasites to total heterotrophic protists and of photosymbionts to total phytoplankton are indicated above each symbol.

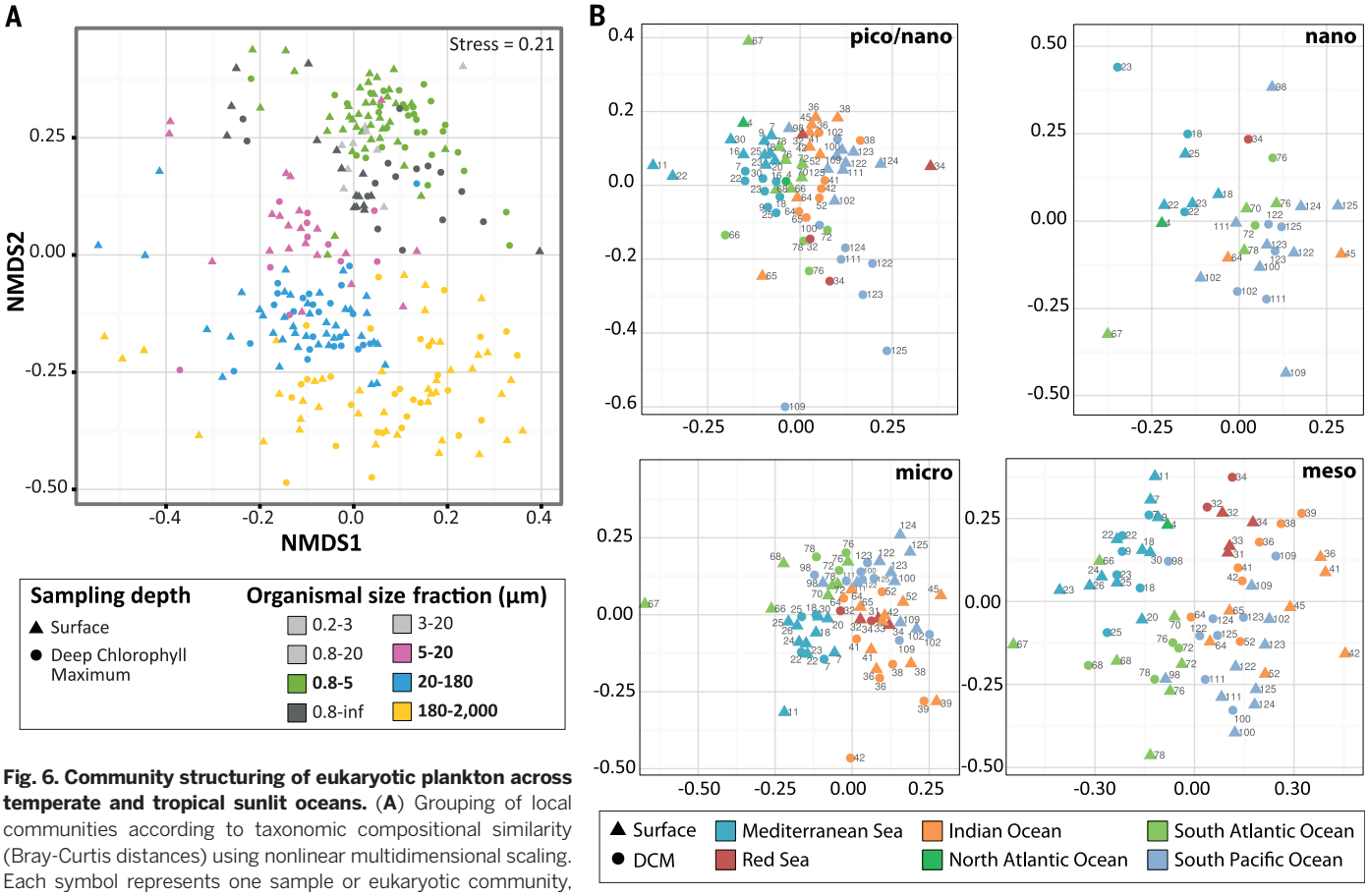


Fig. 6. Community structuring of eukaryotic plankton across temperate and tropical sunlit oceans. (**A**) Grouping of local communities according to taxonomic compositional similarity (Bray-Curtis distances) using nonlinear multidimensional scaling. Each symbol represents one sample or eukaryotic community, corresponding to a particular depth (shape) and organismal size fraction (color). (**B**) Same as in (A), but the different plankton organismal size fractions were analyzed independently, and communities are distinguished by depth (shape) and ocean basins' origin (color). An increasing geographic community differentiation along increasing organismal size fractions is visible and confirmed by the Mantel test [$P = 10^{-3}$, R_m = 0.36, 0.49, 0.50, and 0.51 for the highest piconano- to mesoplankton correlations in Mantel correlograms; see also (*54*)]. In addition, samples from the piconanoplankton only were discriminated by depth (surface versus DCM; $P = 0.001$, $r^2 = 0.2$). The higher diversity and abundance of eukaryotic phototrophs in this fraction (Fig. 5A) may explain overall community structuring by light and, thus, depth.

rare (<0.01%) OTUs were more or less constant across communities, as has been observed in coastal waters (*6*). Only 2 to 17 OTUs (i.e., 0.2 to 8% of total OTUs per and across sample) dominated each community (*54*), suggesting that a small proportion of eukaryotic taxa are key for local plankton ecosystem function. On a worldwide scale, an occurrence-versus-abundance analysis of all ~110,000 *Tara* Oceans OTUs revealed the hyperdominance of cosmopolitan taxa (Fig. 7A). The 381 (0.35% of the total) cosmopolitan OTUs represented ~68% of the total number of reads in the data set. Of these, 269 (71%) OTUs had >100,000 reads and accounted for nearly half (48%) of all rDNA reads (Fig. 7A), a pattern reminiscent of hyperdominance in the largest forest ecosystem on Earth, where only 227 tree species out of an estimated total of 16,000 account for half of all trees in Amazonia (*59*). The cosmopolitan OTUs belonged mainly (314 of 381) to the 11 hyperdiverse eukaryotic planktonic lineages (Fig. 3C) and were essentially phagotrophic (40%) or parasitic (21%), with relatively few (15%) phytoplanktonic taxa (*54*). Of the cosmopolitan OTUs, which represent organisms that are likely among the most abundant eukaryotes on Earth, 25% had poor identity (<95%) to reference taxa, and 11 of these OTUs could not even be affiliated to any available reference sequence (Fig. 7B) (*54*).

Conclusions and perspectives

We used rDNA sequence data to explore the taxonomic and ecological structure of total eukaryotic plankton from the photic oceanic biome, and we integrated these data with existing morphological knowledge. We found that eukaryotic plankton are more diverse than previously thought, especially heterotrophic protists, which may display a wide range of trophic modes (*60*) and include an unsuspected diversity of parasites and photosymbiotic taxa. Dominance of unicellular heterotrophs in plankton ecosystems likely emerged at the dawn of the radiation of eukaryotic cells, together with arguably their most important innovation: phagocytosis. The onset of eukaryophagy in the Neoproterozoic (*61*) probably led to adaptive radiation in heterotrophic eukaryotes through specialization of trophic modes and symbioses, opening novel serial biotic ecological niches. The extensive codiversification of relatively large heterotrophic eukaryotes and their associated parasites supports the idea that biotic interactions, rather than competition for resources and space (*62*), are the primary forces driving organismal diversification in marine plankton systems. Based on rDNA, heterotrophic protists may be even more diverse than prokaryotes in the planktonic ecosystem (*63*). Given that organisms in highly diverse and abundant groups, such as the alveolates and rhizarians, can have genomes more complex than those of humans (*64*), eukaryotic plankton may contain a vast reservoir of unknown marine planktonic genes (*65*). Insights are developing into how heterotrophic protists contribute to a multilayered and integrated ecosystem. The protistan parasites and mutualistic symbionts increase connectivity and complexity of pelagic food webs (*66*, *67*) while contributing to the carbon quota of their larger, longer-lived, and often biomineralized symbiotic hosts, which themselves contribute to carbon export when they die. Decoding the ecological and evolutionary rules governing plankton diversity remains essential for understanding how the

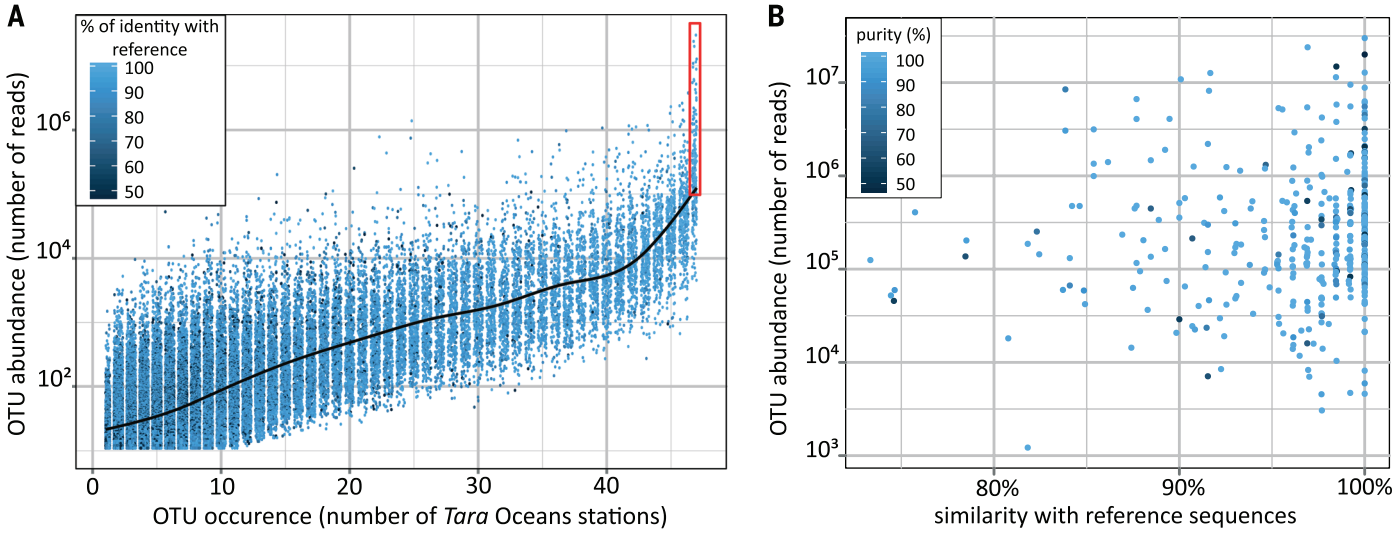


Fig. 7. Cosmopolitanism and abundance of eukaryotic marine plankton. (**A**) Occurrence-versus-abundance plot including the ~110,000 *Tara* Oceans V9 rDNA OTUs. OTUs are colored according to their identity with a reference sequence, and a fitted curve indicates the median OTU size value for each OTU geographic occurrence value. The red rectangle encloses the cosmopolitan and hyperdominant (>10^5 reads) OTUs. (**B**) Similarity to reference barcode and taxonomic purity [a measure of taxonomic assignment consistency defined as the percentage of reads within an OTU assigned to the same taxon; see (*13*)] of the 381 cosmopolitan OTUs, along their abundance (*y* axis).

critical ocean biomes contribute to the functioning of the Earth system.

Materials and methods

V9-18S rDNA for eukaryotic metabarcoding

We used universal eukaryotic primers (*68*) to PCR-amplify (25 cycles in triplicate) the V9-18*S* rDNA genes from all *Tara* Oceans samples. This barcode presents a combination of advantages for addressing general questions of eukaryotic biodiversity over extensive taxonomic and ecological scales: (i) It is universally conserved in length (130 ± 4 base pairs) and simple in secondary structure, thus allowing relatively unbiased PCR amplification across eukaryotic lineages followed by Illumina sequencing. (ii) It includes both stable and highly variable nucleotide positions over evolutionary time frames, allowing discrimination of taxa over a substantial phylogenetic depth. (iii) It is extensively represented in public reference databases across the eukaryotic tree of life, allowing taxonomic assignment among all known eukaryotic lineages (*13*).

Biodiversity analyses

Our bioinformatic pipeline included quality checking (Phred score filtering, elimination of reads without perfect forward and reverse primers, and chimera removal) and conservative filtering (removal of metabarcodes present in less than three reads and two distinct samples). The ~2.3 million metabarcodes (distinct reads) were clustered using an agglomerative, unsupervised single-linkage clustering algorithm, allowing OTUs to reach their natural limits while avoiding arbitrary global clustering thresholds (*13*, *14*). This clustering limited overestimation of biodiversity due to errors in PCR amplification or DNA sequencing, as well as intragenomic polymorphism of rDNA gene copies (*13*). *Tara* Oceans metabarcodes and OTUs were taxonomically assigned by comparison to the 77,449 reference barcodes included in our *V9_PR2* database (*15*). This database derives from the Protist Ribosomal Reference (PR2) database (*69*) but focuses on the V9 region of the gene and includes the following reorganizations: (i) extension of the number of ranks for groups with finer taxonomy (e.g., animals), (ii) expert curation of the taxonomy and renaming in novel environmental groups and dinoflagellates, (iii) resolution of all taxonomic conflicts and inclusion of environmental sequences only if they provide additional phylogenetic information, and (iv) annotation of basic trophic and/or symbiotic modes for all reference barcodes assigned to the genus level [see (*53*) and (*15*) for details]. The *V9_PR2* reference barcodes represent 24,435 species and 13,432 genera from all known major lineages of the tree of eukaryotic life (*15*). Metabarcodes with ≥80% identity to a reference V9 rDNA barcode were considered assignable. Below this threshold it is not possible to discriminate between eukaryotic supergroups, given the short length of V9 rDNA sequences and the relatively fast rate accumulation of substitution mutations in the DNA. In addition to assignment at the finest-possible taxonomic resolution, all assignable metabarcodes were classified into a reference taxonomic framework consisting of 97 major monophyletic groups comprising all known high-rank eukaryotic diversity. This framework, primarily based on a synthesis of protistan biodiversity (*19*), also included all key but still unnamed planktonic clades revealed by previous environmental rDNA clone library surveys (*70*) [e.g., marine alveolates (MALV), marine stramenopiles (MAST), marine ochrophytes (MOCH), and radiolarians (RAD)] (*15*). Details of molecular and bioinformatics methods are available on a companion Web site at http://taraoceans.sb-roscoff.fr/EukDiv/ (*53*). We compiled our data into two databases including the taxonomy, abundance, and size fraction and biogeography information associated with each metabarcode and OTU (*71*).

Ecological inferences

From our *Tara* Oceans metabarcoding data set, we inferred patterns of eukaryotic plankton functional ecology. Based on a literature survey, all reference barcodes assigned to at least the genus level that recruited *Tara* Oceans metabarcodes were associated to basic trophic and symbiotic modes of the organism they come from (*15*) and used for a taxo-functional annotation of our entire metabarcoding data set with the same set of rules used for taxonomic assignation (*53*). False positives were minimized by (i) assigning ecological modes to all individual reference barcodes in *V9_PR2*; (ii) inferring ecological modes to metabarcodes related to monomodal reference barcode(s) (otherwise transferring them to a "NA, nonapplicable" category); and (iii) exploring broad and complex trophic and symbiotic modes that involve fundamental reorganization of the cell structure and metabolism, emerged relatively rarely in the evolutionary history of eukaryotes, and most often concern all known species within monophyletic and ancient groups [see (*15*) for details]. In case of photo- versus heterotrophy, >75% of the major, deep-branching eukaryotic lineages considered (Fig. 3) are monomodal and recruit ~87 and ~69% of all *Tara* Oceans V9 rDNA reads and OTUs, respectively. For parasitism, ~91% of *Tara* Oceans metabarcodes are falling within monophyletic and major groups containing exclusively parasitic species (essentially within the major MALVs groups). Although biases could arise in functional annotation of metabarcodes

relatively distant from reference barcodes in the few complex polymodal groups (e.g., the dinoflagellates that can be phototrophic, heterotrophic, parasitic, or photosymbiotic), a conservative analysis of the trophic and symbiotic ecological patterns presented in Fig. 3, using a ≥99% assignation threshold, shows that these are stable across organismal size fractions and space, independently of the similarity cutoff (80 or 99%), demonstrating their robustness across evolutionary times (*30*).

Note that rDNA gene copy number varies from one to thousands in single eukaryotic genomes (*72*, *73*), precluding direct translation of rDNA read number into abundance of individual organisms. However, the number of rDNA copies per genome correlates positively to the size (*73*) and particularly to the biovolume (*72*) of the eukaryotic cell it represents. We compiled published data from the last ~20 years, confirming the positive correlation between eukaryotic cell size and rDNA copy number across a wide taxonomic and organismal size range [see (*74*); note, however, the ~one order of magnitude of cell size variation for a given rDNA copy number]. To verify whether our molecular ecology protocol preserved this empirical correlation, light microscopy counts of phytoplankton belonging to different eukaryotic supergroups (coccolithophores, diatoms, and dinoflagellates) were performed from nine *Tara* Oceans stations from the Indian, Atlantic, and Southern oceans; transformed into biomass and biovolume data; and then compared with the relative number of V9 rDNA reads found for the identified taxa in the same samples (*74*). Results confirmed the correlation between biovolume and V9 rDNA abundance data ($r^2 = 0.97$, $P = 1 \times 10^{-16}$), although we cannot rule out the possibility that some eukaryotic taxa may not follow the general trend.

REFERENCES AND NOTES

1. C. B. Field, M. J. Behrenfeld, J. T. Randerson, P. Falkowski, Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* **281**, 237–240 (1998). doi: 10.1126/science.281.5374.237; pmid: 9657713
2. D. A. Caron, P. D. Countway, A. C. Jones, D. Y. Kim, A. Schnetzer, Marine protistan diversity. *Annu. Rev. Mar. Sci.* **4**, 467–493 (2012). doi: 10.1146/annurev-marine-120709-142802; pmid: 22457984
3. P. López-García, F. Rodríguez-Valera, C. Pedrós-Alió, D. Moreira, Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409**, 603–607 (2001). doi: 10.1038/35054537; pmid: 11214316
4. S. Y. Moon-van der Staay, R. De Wachter, D. Vaulot, Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409**, 607–610 (2001). doi: 10.1038/35054541; pmid: 11214317
5. B. Díez, C. Pedrós-Alió, R. Massana, Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. *Appl. Environ. Microbiol.* **67**, 2932–2941 (2001). doi: 10.1128/AEM.67.7.2932-2941.2001; pmid: 11425705
6. R. Logares *et al*., Patterns of rare and abundant marine microbial eukaryotes. *Curr. Biol.* **24**, 813–821 (2014). doi: 10.1016/j.cub.2014.02.050; pmid: 24704080
7. V. Edgcomb *et al*., Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. *ISME J.* **5**, 1344–1356 (2011). doi: 10.1038/ismej.2011.6; pmid: 21390079
8. R. Massana *et al*., Phylogenetic and ecological analysis of novel marine stramenopiles. *Appl. Environ. Microbiol.* **70**, 3528–3534 (2004). doi: 10.1128/AEM.70.6.3528-3534.2004; pmid: 15184153
9. L. Guillou *et al*., Widespread occurrence and genetic diversity of marine parasitoids belonging to *Syndiniales* (*Alveolata*). *Environ. Microbiol.* **10**, 3349–3365 (2008). doi: 10.1111/j.1462-2920.2008.01731.x; pmid: 18771501
10. H. Liu *et al*., Extreme diversity in noncalcifying haptophytes explains a major pigment paradox in open oceans. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 12803–12808 (2009). doi: 10.1073/pnas.0905841106; pmid: 19622724
11. Companion Web site: Figure W1 and Database W1 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
12. S. Pesant *et al*., Open science resources for the discovery and analysis of *Tara* Oceans data. http://biorxiv.org/content/early/2015/05/08/019117 (2015).
13. Companion Web site: Text W1 and Figure W2 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
14. F. Mahé, T. Rognes, C. Quince, C. de Vargas, M. Dunthorn, Swarm: Robust and fast clustering method for amplicon-based studies. *PeerJ* **2**, e593 (2014). doi: 10.7717/peerj.593; pmid: 25276506
15. Companion Web site: Database W2, Database W3, and Database W6 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
16. Companion Web site: Text W3, Text W4, Text W5, Figure W4, Figure W5, Figure W6, and Figure W7 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
17. F. W. Preston, The commonness, and rarity, of species. *Ecology* **29**, 254–283 (1948). doi: 10.2307/1930989
18. R. Massana, Eukaryotic picoplankton in surface oceans. *Annu. Rev. Microbiol.* **65**, 91–110 (2011). doi: 10.1146/annurev-micro-090110-102903; pmid: 21639789
19. J. Pawlowski *et al*., CBOL protist working group: Barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLOS Biol.* **10**, e1001419 (2012). doi: 10.1371/journal.pbio.1001419; pmid: 23139639
20. C. Mora, D. P. Tittensor, S. Adl, A. G. B. Simpson, B. Worm, How many species are there on Earth and in the ocean? *PLOS Biol.* **9**, e1001127 (2011). doi: 10.1371/journal.pbio.1001127; pmid: 21886479
21. W. Appeltans *et al*., The magnitude of global marine species diversity. *Curr. Biol.* **22**, 2189–2202 (2012). doi: 10.1016/j.cub.2012.09.036; pmid: 23159596
22. L. M. Márquez, D. J. Miller, J. B. MacKenzie, M. J. H. Van Oppen, Pseudogenes contribute to the extreme diversity of nuclear ribosomal DNA in the hard coral *Acropora*. *Mol. Biol. Evol.* **20**, 1077–1086 (2003). doi: 10.1093/molbev/msg122; pmid: 12777522
23. S. R. Santos, R. A. Kinzie III, K. Sakai, M. A. Coffroth, Molecular characterization of nuclear small subunit (18S)-rDNA pseudogenes in a symbiotic dinoflagellate (*Symbiodinium*, Dinophyta). *J. Eukaryot. Microbiol.* **50**, 417–421 (2003). doi: 10.1111/j.1550-7408.2003.tb00264.x; pmid: 14733432
24. J. Decelle, S. Romac, E. Sasaki, F. Not, F. Mahé, Intracellular diversity of the V4 and V9 regions of the 18S rRNA in marine protists (radiolarians) assessed by high-throughput sequencing. *PLOS ONE* **9**, e104297 (2014). doi: 10.1371/journal.pone.0104297; pmid: 25090095
25. A. Sournia, M.-J. Chrétiennot-Dinet, M. Ricard, Marine phytoplankton: How many species in the world ocean? *J. Plankton Res.* **13**, 1093–1099 (1991). doi: 10.1093/plankt/13.5.1093
26. P. H. Wiebe *et al*., Deep-sea sampling on CMarZ cruises in the Atlantic Ocean – An introduction. *Deep-Sea Res. Part II* **57**, 2157–2166 (2010). doi: 10.1016/j.dsr2.2010.09.018
27. D. Boltovskoy, Diversity and endemism in cold waters of the South Atlantic: Contrasting patterns in the plankton and the benthos. *Sci. Mar.* **69**, 17–26 (2005).
28. C. de Vargas, R. Norris, L. Zaninetti, S. W. Gibb, J. Pawlowski, Molecular evidence of cryptic speciation in planktonic foraminifers and their relation to oceanic provinces. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 2864–2868 (1999). doi: 10.1073/pnas.96.6.2864; pmid: 10077602
29. K. M. K. Halbert, E. Goetze, D. B. Carlon, High cryptic diversity across the global range of the migratory planktonic copepods *Pleuromamma piseki* and *P. gracilis*. *PLOS ONE* **8**, e77011 (2013). doi: 10.1371/journal.pone.0077011; pmid: 24167556
30. Companion Web site: Figure W8, Figure W9, Figure W10, and Figure W14 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
31. F. Burki, P. J. Keeling, Rhizaria. *Curr. Biol.* **24**, R103–R107 (2014). doi: 10.1016/j.cub.2013.12.025; pmid: 24502779
32. N. R. Swanberg, thesis, Massachussetts Institute of Technology (1974).
33. I. Probert *et al*., *Brandtodinium* gen. nov. and *B. nutricula* comb. Nov. (Dinophyceae), a dinoflagellate commonly found in symbiosis with polycystine radiolarians. *J. Phycol.* **50**, 388–399 (2014). doi: 10.1111/jpy.12174
34. S. von der Heyden, E. E. Chao, K. Vickerman, T. Cavalier-Smith, Ribosomal RNA phylogeny of bodonid and diplonemid flagellates and the evolution of euglenozoa. *J. Eukaryot. Microbiol.* **51**, 402–416 (2004). doi: 10.1111/j.1550-7408.2004.tb00387.x; pmid: 15352322
35. E. Schnepf, Light and electron microscopical observations in *Rhynchopus coscinodiscivorus* spec. nov., a Colorless, phagotrophic euglenozoon with concealed flagella. *Arch. Protistenkd.* **144**, 63–74 (1994). doi: 10.1016/S0003-9365(11)80225-3
36. M. R. Dennett, Video plankton recorder reveals high abundances of colonial Radiolaria in surface waters of the central North Pacific. *J. Plankton Res.* **24**, 797–805 (2002). doi: 10.1093/plankt/24.8.797
37. L. Stemmann *et al*., Global zoogeography of fragile macrozooplankton in the upper 100-1000 m inferred from the underwater video profiler. *ICES J. Mar. Sci.* **65**, 433–442 (2008). doi: 10.1093/icesjms/fsn010
38. A. F. Michaels, D. A. Caron, N. R. Swanberg, F. A. Howse, C. M. Michaels, Planktonic sarcodines (Acantharia, Radiolaria, Foraminifera) in surface waters near Bermuda: Abundance, biomass and vertical flux. *J. Plankton Res.* **17**, 131–163 (1995). doi: 10.1093/plankt/17.1.131
39. E. Haeckel, "Report on the Radiolaria collected by H.M.S. Challenger during the years 1873-1876" in *Report on the Scientific Results of the Voyage of* H.M.S. Challenger *During the Years 1873-76. Zoology.* (Neill, Edinburgh, 1887).
40. R. Siano, M. Montresor, I. Probert, F. Not, C. de Vargas, *Pelagodinium* gen. nov. and *P. béii* comb. nov., a dinoflagellate symbiont of planktonic foraminifera. *Protist* **161**, 385–399 (2010). doi: 10.1016/j.protis.2010.01.002; pmid: 20149979
41. J. Decelle *et al*., An original mode of symbiosis in open ocean plankton. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 18000–18005 (2012). doi: 10.1073/pnas.1212303109; pmid: 23071304
42. Y. Shaked, C. de Vargas, Pelagic photosymbiosis: rDNA assessment of diversity and evolution of dinoflagellate symbionts and planktonic foraminiferal hosts. *Mar. Ecol. Prog. Ser.* **325**, 59–71 (2006). doi: 10.3354/meps325059
43. J. Decelle, New perspectives on the functioning and evolution of photosymbiosis in plankton: Mutualism or parasitism? *Commun. Integr. Biol.* **6**, e24560 (2013). doi: 10.4161/cib.24560; pmid: 23986805
44. R. Siano *et al*., Distribution and host diversity of Amoebophryidae parasites across oligotrophic waters of the Mediterranean Sea. *Biogeosciences* **8**, 267–278 (2011). doi: 10.5194/bg-8-267-2011
45. D. Coats, M. Park, Parasitism of photosynthetic dinoflagellates by three strains of *Amoebophyra* (Dinophyta): Parasite survival, infectivity, generation time, and host specificity. *J. Phycol.* **528**, 520–528 (2002). doi: 10.1046/j.1529-8817.2002.01200.x
46. K. E. Wommack, R. R. Colwell, Virioplankton: Viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **64**, 69–114 (2000). doi: 10.1128/MMBR.64.1.69-114.2000; pmid: 10704475
47. A. Skovgaard, Dirty tricks in the plankton: Diversity and role of marine parasitic protists. *Acta Protozool.* **53**, 51–62 (2014).
48. J. Bråte *et al*., Radiolaria associated with large diversity of marine alveolates. *Protist* **163**, 767–777 (2012). doi: 10.1016/j.protis.2012.04.004; pmid: 22658831
49. T. R. Bachvaroff, S. Kim, L. Guillou, C. F. Delwiche, D. W. Coats, Molecular diversity of the syndinean genus *Eudubosquella* based on single-cell PCR analysis. *Appl. Environ. Microbiol.* **78**, 334–345 (2012). doi: 10.1128/AEM.06678-11; pmid: 22081578
50. S. Rueckert, T. G. Simdyanov, V. V. Aleoshin, B. S. Leander, Identification of a divergent environmental DNA sequence clade using the phylogeny of gregarine parasites (Apicomplexa) from crustacean hosts. *PLOS ONE* **6**, e18163 (2011). doi: 10.1371/journal.pone.0018163; pmid: 21483868
51. A. Skovgaard, S. A. Karpov, L. Guillou, The parasitic dinoflagellates *Blastodinium* spp. inhabiting the gut of marine, planktonic copepods: Morphology, ecology, and unrecognized species diversity. *Front. Microbiol.* **3**, 305 (2012). doi: 10.3389/fmicb.2012.00305; pmid: 22973263
52. H. McCallum *et al*., Does terrestrial epidemiology apply to marine systems? *Trends Ecol. Evol.* **19**, 585–591 (2004). doi: 10.1016/j.tree.2004.08.009

53. Companion Web site: detailed Material and Methods, Database W9, and Figure W11 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
54. Companion Web site: Figure W12, Figure W13, Database W7, and Database W8 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
55. M. Holyoak, M. A. Leibold, R. D. Holt, Eds., *Metacommunities: Spatial Dynamics and Ecological Communities* (University of Chicago Press, Chicago, 2005).
56. L. G. M. Baas Becking, *Geobiologie of Inleiding tot de Milieukunde* (W. P. Van Stockum and Zoon, The Hague, Netherlands, 1934).
57. K. T. C. A. Peijnenburg, E. Goetze, High evolutionary potential of marine zooplankton. *Ecol. Evol.* **3**, 2765–2781 (2013). doi: 10.1002/ece3.644; pmid: 24567838
58. V. Smetacek, Microbial food webs. The ocean's veil. *Nature* **419**, 565 (2002). doi: 10.1038/419565a; pmid: 12374956
59. H. ter Steege *et al*., Hyperdominance in the Amazonian tree flora. *Science* **342**, 1243092 (2013). doi: 10.1126/science.1243092; pmid: 24136971
60. D. Vaulot, K. Romari, F. Not, Are autotrophs less diverse than heterotrophs in marine picoplankton? **10**, 266–267 (2002).
61. A. H. Knoll, Paleobiological perspectives on early eukaryotic evolution. *Cold Spring Harb. Perspect. Biol.* **6**, 1–14 (2014). doi: 10.1101/cshperspect.a016121; pmid: 24384569
62. V. Smetacek, A watery arms race. *Nature* **411**, 745 (2001). doi: 10.1038/35081210; pmid: 11459035
63. S. Sunagawa *et al*., Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
64. M. J. Oliver, D. Petrov, D. Ackerly, P. Falkowski, O. M. Schofield, The mode and tempo of genome size evolution in eukaryotes. *Genome Res.* **17**, 594–601 (2007). doi: 10.1101/gr.6096207; pmid: 17420184
65. H. Abida *et al*., Bioprospecting marine plankton. *Mar. Drugs* **11**, 4594–4611 (2013). doi: 10.3390/md11114594; pmid: 24240981
66. G. Lima-Mendez *et al*., Determinants of community structure in the global plankton interactome. *Science* **348**, 1262073 (2015).
67. K. D. Lafferty, A. P. Dobson, A. M. Kuris, Parasites dominate food web links. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 11211–11216 (2006). doi: 10.1073/pnas.0604755103; pmid: 16844774
68. L. A. Amaral-Zettler, E. A. McCliment, H. W. Ducklow, S. M. Huse, A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLOS ONE* **4**, e6372 (2009). doi: 10.1371/journal.pone.0006372; pmid: 19633714
69. L. Guillou *et al*., The Protist Ribosomal Reference database (PR^2): A catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* **41**, D597–D604 (2013). doi: 10.1093/nar/gks1160; pmid: 23193267
70. R. Massana, J. del Campo, M. E. Sieracki, S. Audic, R. Logares, Exploring the uncultured microeukaryote majority in the oceans: Reevaluation of ribogroups within stramenopiles. *ISME J.* **8**, 854–866 (2014). doi: 10.1038/ismej.2013.204; pmid: 24196325
71. Companion Web site: Database W4 and Database W5 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
72. A. Godhe *et al*., Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. *Appl. Environ. Microbiol.* **74**, 7174–7182 (2008). doi: 10.1128/AEM.01298-08; pmid: 18849462
73. F. Zhu, R. Massana, F. Not, D. Marie, D. Vaulot, Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol. Ecol.* **52**, 79–92 (2005). doi: 10.1016/j.femsec.2004.10.006; pmid: 16329895
74. Companion Web site: Text W2 and Figure W3 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
75. M. Kim, S. Nam, W. Shin, D. W. Coats, M. Park, *Dinophysis caudata* (dinophyceae) sequesters and retains plastids from the mixotrophic ciliate prey *Mesodinium rubrum*. *J. Phycol.* **48**, 569–579 (2012). doi: 10.1111/j.1529-8817.2012.01150.x

ACKNOWLEDGMENTS

We thank the following people and sponsors for their commitment: CNRS (in particular, the GDR3280); EMBL; Genoscope/CEA; UPMC; VIB; Stazione Zoologica Anton Dohrn; UNIMIB; Rega Institute; KU Leuven; Fund for Scientific Research – The French Ministry of Research, the French Government "Investissements d'Avenir" programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), and MEMO LIFE (ANR-10-LABX-54); PSL* Research University (ANR-11-IDEX-0001-02); ANR (projects POSEIDON/ANR-09-BLAN-0348, PROMETHEUS/ANR-09-PCS-GENM-217, PHYTBACK/ANR-2010-1709-01, and TARA-GIRUS/ANR-09-PCS-GENM-218); EU FP7 (MicroB3/No.287589, IHMS/HEALTH-F4-2010-261376); European Research Council Advanced Grant Awards to C. Bowler (Diatomite:294823); Gordon and Betty Moore Foundation grant 3790 to M.B.S.; Spanish Ministry of Science and Innovation grant CGL2011-26848/BOS MicroOcean PANGENOMICS and TANIT (CONES 2010-0036) grant from the Agency for Administration of University and Research Grants (AGAUR) to S.G.A.; and Japan Society for the Promotion of Science KAKENHI grant 26430184 to H.O. We also thank the following for their support and commitment: A. Bourgois, E. Bourgois, R. Troublé, Région Bretagne, G. Ricono, the Veolia Environment Foundation, Lorient Agglomération, World Courier, Illumina, the Electricité de France Foundation, Fondation pour la Recherche sur la Biodiversité, the Prince Albert II de Monaco Foundation, and the *Tara* schooner and its captains and crew. We thank MERCATOR-CORIOLIS and ACRI-ST for providing daily satellite data during the expedition. We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries who granted sampling permissions. *Tara* Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org). We also acknowledge assistance from European Bioinformatics Institute (EBI) (in particular, G. Cochrane and P. ten Hoopen) as well as the EMBL Advanced Light Microscopy Facility (in particular, R. Pepperkok). We thank F. Gaill, B. Kloareg, F. Lallier, D. Boltovskoy, A. Knoll, D. Richter, and E. Médard for help and advice on the manuscript. We declare that all data reported herein are fully and freely available from the date of publication, with no restrictions, and that all of the samples, analyses, publications, and ownership of data are free from legal entanglement or restriction of any sort by the various nations whose waters the *Tara* Oceans expedition sampled in. Data described herein are available at http://taraoceans.sb-roscoff.fr/EukDiv/, at EBI under the project IDs PRJEB402 and PRJEB6610, and at PANGAEA (see table S1). The data release policy regarding future public release of *Tara* Oceans data is described in (*12*). All authors approved the final manuscript. This article is contribution number 24 of *Tara* Oceans. The supplementary materials contain additional data.

Tara Oceans Coordinators

Silvia G. Acinas,[1] Peer Bork,[2] Emmanuel Boss,[3] Chris Bowler,[4] Colomban de Vargas,[5,6] Michael Follows,[7] Gabriel Gorsky,[8,9] Nigel Grimsley,[10,11] Pascal Hingamp,[12] Daniele Iudicone,[13] Olivier Jaillon,[14,15,16] Stefanie Kandels-Lewis,[2,17] Lee Karp-Boss,[3] Eric Karsenti,[4,17] Uros Krzic,[18] Fabrice Not,[5,6] Hiroyuki Ogata,[19] Stephane Pesant,[20,21] Jeroen Raes,[22,23,24] Emmanuel G. Reynaud,[25] Christian Sardet,[26,27] Mike Sieracki,[28] Sabrina Speich,[29,30] Lars Stemmann,[8] Matthew B. Sullivan,[31]* Shinichi Sunagawa,[2] Didier Velayoudon,[32] Jean Weissenbach,[14,15,16] Patrick Wincker[14,15,16]

[1]Department of Marine Biology and Oceanography, ICM-CSIC, Passeig Marítim de la Barceloneta, 37-49, Barcelona E08003, Spain. [2]Structural and Computational Biology, EMBL, Meyerhofstraße 1, 69117 Heidelberg, Germany. [3]School of Marine Sciences, University of Maine, Orono, ME 04469, USA. [4]Ecole Normale Supérieure, IBENS, and Inserm U1024, and CNRS UMR 8197, Paris, F-75005 France. [5]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [6]Sorbonne Universités, UPMC Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [7]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [8]CNRS, UMR 7093, LOV, Observatoire Océanologique, F-06230, Villefranche-sur-Mer, France. [9]Sorbonne Universités, UPMC Paris 06, UMR 7093, LOV, Observatoire Océanologique, F-06230, Villefranche-sur-Mer, France. [10]CNRS UMR 7232, BIOM, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [11]Sorbonne Universités Paris 06, OOB UPMC, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [12]Aix Marseille Université, CNRS IGS, UMR 7256, 13288 Marseille, France. [13]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [14]CEA, Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France. [15]CNRS, UMR 8030, CP5706 Evry, France. [16]Université d'Evry, UMR 8030, CP5706 Evry, France. [17]Directors' Research, EMBL, Meyerhofstraße 1, 69117 Heidelberg, Germany. [18]Cell Biology and Biophysics, EMBL, Meyerhofstraße 1, 69117 Heidelberg, Germany. [19]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan. [20]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany. [21]MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany. [22]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [23]Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. [24]Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [25]Earth Institute, University College Dublin, Dublin, Ireland. [26]CNRS, UMR 7009 Biodev, Observatoire Océanologique, F-06230 Villefranche-sur-Mer, France. [27]Sorbonne Universités, UPMC Univ Paris 06, UMR 7009 Biodev, F-06230 Observatoire Océanologique, Villefranche-sur-Mer, France. [28]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME 04544, USA. [29]Department of Geosciences, LMD, Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris, Cedex 05, France. [30]Laboratoire de Physique des Océan UBO-IUEM Place Copernic 29820 Plouzané, France. [31]Department of Ecology and Evolutionary Biology, University of Arizona, 1007 East Lowell Street, Tucson, AZ 85721, USA. [32]DVIP Consulting, Sèvres, France.
*Present address: Department of Microbiology, Ohio State University, Columbus, OH 43210, USA.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/1261605/suppl/DC1
Table S1
Appendix S1

23 September 2014; accepted 27 February 2015
10.1126/science.1261605

OCEAN PLANKTON

Environmental characteristics of Agulhas rings affect interocean plankton transport

Emilie Villar,[1]* Gregory K. Farrant,[2,3]† Michael Follows,[11]† Laurence Garczarek,[2,3]† Sabrina Speich,[5,23]¶ Stéphane Audic,[2,3] Lucie Bittner,[2,3,4]‡ Bruno Blanke,[5] Jennifer R. Brum,[6]** Christophe Brunet,[7] Raffaella Casotti,[7] Alison Chase,[8] John R. Dolan,[9,10] Fabrizio d'Ortenzio,[9,10] Jean-Pierre Gattuso,[9,10] Nicolas Grima,[5] Lionel Guidi,[9,10] Christopher N. Hill,[11] Oliver Jahn,[11] Jean-Louis Jamet,[12] Hervé Le Goff,[13] Cyrille Lepoivre,[1] Shruti Malviya,[4] Eric Pelletier,[14,15,16] Jean-Baptiste Romagnan,[9,10] Simon Roux,[6]** Sébastien Santini,[1] Eleonora Scalco,[7] Sarah M. Schwenck,[6] Atsuko Tanaka,[4]§ Pierre Testor,[13] Thomas Vannier,[14,15,16] Flora Vincent,[4] Adriana Zingone,[7] Céline Dimier,[2,3,4] Marc Picheral,[9,10] Sarah Searson,[9,10]|| Stefanie Kandels-Lewis,[17,18] *Tara* Oceans Coordinators¶ Silvia G. Acinas,[19] Peer Bork,[17,20] Emmanuel Boss,[8] Colomban de Vargas,[2,3] Gabriel Gorsky,[9,10] Hiroyuki Ogata,[1]# Stéphane Pesant,[21,22] Matthew B. Sullivan,[6]** Shinichi Sunagawa,[17] Patrick Wincker,[14,15,16] Eric Karsenti,[4,18]* Chris Bowler,[4]* Fabrice Not,[2,3]*†† Pascal Hingamp,[1]*†† Daniele Iudicone[7]*††

Agulhas rings provide the principal route for ocean waters to circulate from the Indo-Pacific to the Atlantic basin. Their influence on global ocean circulation is well known, but their role in plankton transport is largely unexplored. We show that, although the coarse taxonomic structure of plankton communities is continuous across the Agulhas choke point, South Atlantic plankton diversity is altered compared with Indian Ocean source populations. Modeling and in situ sampling of a young Agulhas ring indicate that strong vertical mixing drives complex nitrogen cycling, shaping community metabolism and biogeochemical signatures as the ring and associated plankton transit westward. The peculiar local environment inside Agulhas rings may provide a selective mechanism contributing to the limited dispersal of Indian Ocean plankton populations into the Atlantic.

The Agulhas Current, which flows down the east coast of Africa, leaks from the Indo-Pacific Ocean into the Atlantic Ocean (*1*). This leakage, a choke point to heat and salt distribution across the world's oceans, has been increasing over the last decades (*2*). The influence of the Agulhas leakage on global oceanic circulation makes this area a sensitive lever in climate change scenarios (*3*). Agulhas leakage has been a gateway for planetary-scale water transport since the early Pleistocene (*4*), but diatom fossil records suggest that it is not a barrier to plankton dispersal (*5*). Most of the Agulhas leakage occurs through huge anticyclonic eddies known as Agulhas rings. These 100- to 400-km-diameter rings bud from Indian Ocean subtropical waters at the Agulhas Retroflection (*1*). Each year, up to half a dozen Agulhas rings escape the Indian Ocean, enter Cape Basin, and drift northwesterly across the South Atlantic, reaching the South American continent over the course of several years (*1*, *6*). During the transit of Agulhas rings, strong westerly "roaring forties" winds prevalent in the southern 40s and 50s latitudes cause intense internal cooling and mixing (*7*).

We studied the effect of Agulhas rings and the environmental changes they sustain on plankton dispersal. Plankton such as microalgae, which produce half of the atmospheric oxygen derived from photosynthesis each year, are at the base of open-ocean ecosystem food chains, thus playing an essential role in the functioning of the biosphere. Their dispersal is critical for marine ecosystem resilience in the face of environmental change (*8*). As part of the *Tara* Oceans expedition (*9*), we describe taxonomic and functional plankton assemblages inside Agulhas rings and across the three oceanic systems that converge at the Agulhas choke point: the western Indian Ocean subtropical gyre, the South Atlantic Ocean gyre, and the Southern Ocean below the Antarctic Circumpolar Current (Fig. 1).

Physical and biological oceanography of the sampling sites

The Indian, South Atlantic, and Southern Oceans were each represented by three sites sampled between May 2010 and January 2011 (Fig. 1 and table S1). A wide range of environmental conditions were encountered (*10*). We first sampled the two large contiguous Indian and South Atlantic subtropical gyres and the Agulhas ring structures that maintain the physical connection between them. On the western side of the Indian Ocean, station TARA_052 was characterized by tropical, oligotrophic conditions. Station TARA_064 was located within an anticyclonic eddy representing the Agulhas Current recirculation. Station TARA_065 was located at the inner edge of the Agulhas Current on the South African slope that feeds the Agulhas retroflection and Agulhas ring formation (*3*). In the South Atlantic Ocean, station TARA_070, sampled in late winter, was located in the eastern subtropical Atlantic basin. Station TARA_072 was located within the tropical circulation of the South Atlantic Ocean, and Station TARA_076 was at the northwest extreme of the South Atlantic subtropical gyre. Two stations (TARA_068 and TARA_078) from the west and east South Atlantic Ocean sampled Agulhas rings. Three stations (TARA_082, TARA_084, and TARA_085) in the Southern Ocean were selected to sample the Antarctic Circumpolar Current frontal system. Station TARA_082 sampled sub-Antarctic waters flowing northward along the Argentinian slope, waters that flow along the Antarctic Circumpolar Current (*11*) with characteristics typical

[1]Aix Marseille Université, CNRS, IGS UMR 7256, 13288 Marseille, France. [2]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [3]Sorbonne Universités, Université Pierre et Marie Curie UPMC, Université Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [4]Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), and Inserm U1024, and CNRS UMR 8197, F-75005 Paris, France. [5]Laboratoire de Physique des Océans (LPO) UMR 6523 CNRS-Ifremer-IRD-UBO, Plouzané, France. [6]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA. [7]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [8]School of Marine Sciences, University of Maine, Orono, ME, USA. [9]Sorbonne Universités, UPMC Université Paris 06, Observatoire Océanologique, F-06230 Villefranche-sur-Mer, France. [10]INSU-CNRS, UMR 7093, LOV, Observatoire Océanologique, F-06230 Villefranche-sur-Mer, France. [11]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [12]Université de Toulon, Laboratoire PROTEE-EBMA E.A. 3819, BP 20132, 83957 La Garde Cedex, France. [13]CNRS, UMR 7159, Laboratoire d'Océanographie et du Climat LOCEAN, 4 Place Jussieu, 75005 Paris, France. [14]Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), Institut de Génomique, Genoscope, 2 Rue Gaston Crémieux, 91057 Evry, France. [15]CNRS, UMR 8030, CP5706, Evry, France. [16]Université d'Evry, UMR 8030, CP5706, Evry, France. [17]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [18]Directors' Research, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [19]Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM), CSIC, Passeig Marítim de la Barceloneta, 37-49, Barcelona E08003, Spain. [20]Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany. [21]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany. [22]MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany. [23]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD) UMR 8539, Ecole Normale Supérieure, 24 Rue Lhomond, 75231 Paris Cedex 05, France. ***Corresponding author. E-mail: villar@igs.cnrs-mrs.fr (E.V.); not@sb-roscoff.fr (F.N.); hingamp@igs.cnrs-mrs.fr (P.H.); iudicone@szn.it (D.I.); karsenti@embl.de (E.K.), cbowler@biologie.ens.fr (C.B.)** †These authors contributed equally to this work. ‡Present address: CNRS FR3631, Institut de Biologie Paris-Seine, F-75005 Paris, France; Sorbonne Universités, UPMC Université Paris 06, Institut de Biologie Paris-Seine (IBPS), F-75005 Paris, France. §Present address: Muroran Marine Station, Field Science Center for Northern Biosphere, Hokkaido University, Japan. ||Present address: CMORE, University Hawaii, Honolulu, USA. ¶*Tara* Oceans coordinators are listed at the end of this paper. #Present address: Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan. **Present address: Department of Microbiology, Ohio State University, Columbus, OH 43210, USA. ††These authors contributed equally to this work.

of summer sub-Antarctic surface waters and are stratified by seasonal heating. Station TARA_084 was located on the southern part of the Antarctic Circumpolar Current, in the Drake Passage between the Polar Front and the South Antarctic Circumpolar Current front (*11*). Station TARA_085 was located on the southern edge of the South Antarctic Circumpolar Current front with waters typical of polar regions.

We compared overall plankton community structures between the three oceans using imaging and genetic surveys of samples from the epipelagic zone of each station (*12*). Prokaryote, phyto-, and zooplankton assemblages were similar across Indian and South Atlantic Ocean samples but different from Southern Ocean samples (Fig. 2A). In the Indian and South Atlantic Oceans, zooplankton communities were dominated by Calanoida, Cyclopoida (Oithonidae), and Poecilostomatoida copepods (*12*); phytoplankton communities were mainly composed of chlorophytes, pelagophytes, and haptophytes (*12*). In contrast, Southern Ocean zooplankton communities were distinguished by an abundance of *Limacina* spp. gastropods and Poecilostomatoida copepods. Southern Ocean phytoplankton were primarily diatoms and haptophytes. The divergence was even more conspicuous with respect to prokaryotes, in that picocyanobacteria, dominant in the Indian and South Atlantic Oceans, were absent in the Southern Ocean. The Southern Ocean had a high proportion of Flavobacteria and Rhodobacterales (*12*). Virus concentrations in the <0.2-μm size fractions were significantly lower in the southernmost Southern Ocean station (*13*). Viral particles were significantly smaller in two of the three Southern Ocean sampling sites, and two Southern Ocean viromes had significantly lower richness compared with the South Atlantic and Indian Oceans (*13*). Although nucleocytoplasmic large DNA viruses were similarly distributed in the South Atlantic and Indian Oceans (*12*), two Southern Ocean sites contained coccolithoviruses also found in the TARA_068 Agulhas ring but not in the other Indian and South Atlantic stations.

Biological connection across the Agulhas choke point

Genetic material as represented by ribosomal RNA gene (rDNA) sequences showed exchange patterns across the oceans (shared barcode richness) (*14*). Despite a smaller interface between the Indian and South Atlantic Oceans than either have with the Southern Ocean, more than three times as much genetic material was in common between the Indian and South Atlantic Oceans than either had with the Southern Ocean (Fig. 2B) (*15*). Indeed, the Indian–South Atlantic interocean shared barcodes richness (32 ± 5%) was not significantly different from typical intraocean values (37 ± 7%, Tukey post hoc, 0.95 confidence). Shared barcode richness involving the Southern Ocean was significantly lower (9 ± 3%) (Fig. 2C). We found that the proportion of whole shotgun metagenomic reads shared between samples, both intraoceanic and Indian–South Atlantic interocean similarities, were in the 18 to 30% range, whereas interocean similarities with Southern Ocean samples were only 5 to 6% (*16*). The statistically indistinguishable Indo-Atlantic intra- and interocean genetic similarities revealed a high Indo-Atlantic biological connection despite the physical basin discontinuity.

Nonetheless, differences on either side of the Agulhas choke point were evident. We found that prokaryote barcode richness was greater in the South Atlantic than in the Indian Ocean (Fig. 3A) (0.2- to 3-μm size fraction). The opposite trend characterized eukaryotes larger than 20 μm in size. We cannot rule out the possibility that the higher prokaryote diversity observed in the South Atlantic Ocean might be due to a protocol artifact resulting from a difference in prefiltration pore size from 1.6 μm (Indian Ocean) to 3 μm (South Atlantic and Southern Oceans). As also evident from the panoceanic *Tara* Oceans data set (*17*), smaller size fractions showed greater eukaryote diversity across the Agulhas system. In all size fractions that we analyzed, samples from the Southern Ocean were less diverse than samples from the South Atlantic Ocean and Indian Ocean (Fig. 3A).

When rDNA barcodes were clustered by sequence similarity and considered at operational taxonomic unit (OTU) level (*14*), more than half (57%) of the OTUs contained higher sub-OTU barcode richness in the Indian Ocean than in the South Atlantic Ocean, whereas less than a third (32%) of OTUs were richer in the South Atlantic Ocean, leaving only 11% as strictly cosmopolitan (Fig. 3B). Taken together, these 1307 OTUs represented 98% of the barcode abundance, indicating that the observed higher barcode richness within OTUs in the Indian Ocean was not conferred by the rare biosphere. Certain taxa displayed unusual sub-OTU richness profiles across the choke point. Consistent with their relatively large size, Opisthokonta (mostly copepods), Rhizaria (such as radiolarians), and Stramenopiles (in particular diatoms) had much higher sub-OTU barcode richness in the Indian Ocean, whereas only small-sized Hacrobia (mostly haptophytes) showed modest increased sub-OTU barcode richness in the South Atlantic Ocean. The plankton filtering that we observed in fractions above 20 μm through the Agulhas choke point might explain the reduction of marine nekton diversity from the Indian Ocean to the South Atlantic Ocean (*18*) by propagating up the food web (*19*).

In situ sampling of two Agulhas rings

To understand whether the environment of Agulhas rings, the main transporters of water across the choke point, might act as a biological filter between the Indian Ocean and the South Atlantic Ocean, we analyzed data collected in both a young and an old Agulhas ring. The young ring sampled at station TARA_068 was located in the Cape Basin, west of South Africa, where rings are often observed after their formation at the Agulhas Retroflection (*7*, *20*). It was a large Agulhas ring that detached from the retroflection about 9 to 10 months before sampling. This ring first moved northward and then westward in the Cape Basin while interacting with other structures (red track in Fig. 1) (*21*). Ocean color data collected by satellite showed that surface chlorophyll concentrations were higher in the Cape Basin than at the retroflection, suggesting that vigorous vertical

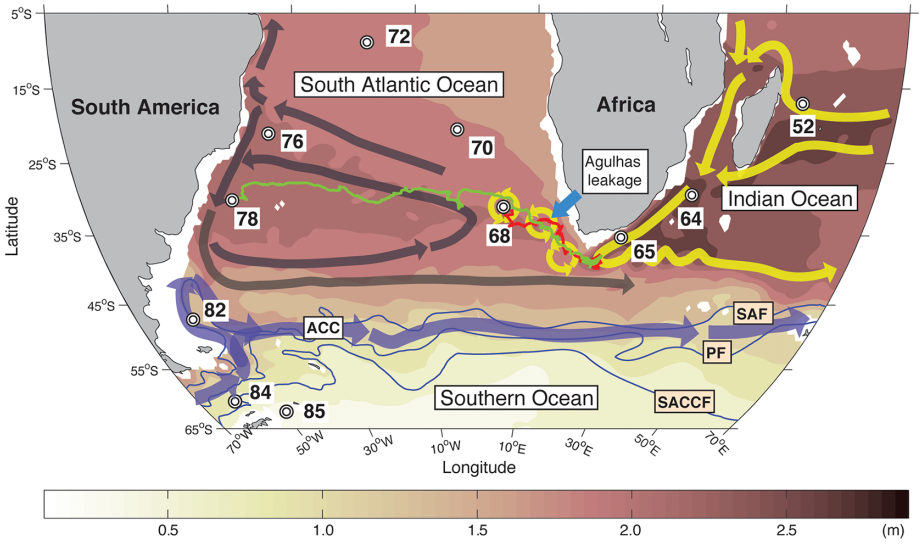


Fig. 1. The oceanic circulation around the Agulhas choke point and location of *Tara* Oceans stations. The map shows the location of sampling stations, together with trajectories of the young and old Agulhas rings (TARA_068 and TARA_078, red and green tracks, respectively). The stations here considered as representative of the main basins are (i) TARA_052, TARA_064, and TARA_065 for Indian Ocean; (ii) TARA_070, TARA_072, and TARA_076 for the South Atlantic Ocean, and (iii) TARA_082, TARA_084, and TARA_085 for the Southern Ocean. The mean ocean circulation is schematized by arrows (currents) and background colors [surface climatological dynamic height (0/2000 dbar from CARS2009; www.cmar.csiro.au/cars)] (*70*). Agulhas rings are depicted as circles. The Antarctic Circumpolar Current front positions are from (*13*).

mixing might have occurred in the Cape Basin (*22*). At the time of sampling, the anticyclonic Agulhas ring was 130 to 150 km in diameter, was about 30 cm higher than average sea surface height, and was flanked by a 130- to 150-km cyclonic eddy to the north and a larger (>200 km) one to the east (Fig. 4A) (*23*). Thermosalinograph data showed that filaments of colder, fresher water surrounded the young ring core (Fig. 4A) (*23*). To position the biological sampling station close to the ring core, a series of conductivity-temperature-depth (CTD) casts was performed (*23*, *24*). The young Agulhas ring had a surface temperature and salinity of 16.8°C and 35.7 practical salinity units (PSU), respectively, and the isopycnal sloping could be traced down to CTD maximal depth (900 to 1000 m). The core of the ring water was 5°C cooler than Indian Ocean subtropical source waters at similar latitudes (TARA_065) (table S1), typical for the subtropical waters south of Africa (17.8°C, 35.56 PSU, respectively) (*25*). The mixed layer of the young ring was deep (>250 m) compared with seasonal cycles of the mixed layer depths in the region (50 to 100 m) (Fig. 4C), typical of Agulhas rings (*26*). At larger scales (Fig. 4B) (*24*), steep spatial gradients were observed, with fresher and colder water in the Cape Basin than in the Agulhas Current because of both lateral mixing with waters from the south and surface fluxes. This confirms that the low temperature of the young Agulhas ring is a general feature of this Indian to South Atlantic Ocean transitional basin. Air-sea exchanges of heat and momentum promoted convection in the ring core, which was not compensated by lateral mixing and advection. The core of the Agulhas ring thus behaved as a subpolar environment traveling across a subtropical region.

At station TARA_078, we sampled a second structure whose origins were in the Agulhas Retroflection, likely a 3-year-old Agulhas ring. This old ring, having crossed the South Atlantic Ocean, was being absorbed by the western boundary current of the South Atlantic subtropical gyre. The structure sampled at station TARA_078 was characterized by a warm salty core (*27*). As for the young Agulhas ring sampled, the old ring also had a 100-m-deeper pycnocline than surrounding waters, typical of large anticyclonic structures.

The plankton assemblage of both Agulhas rings most closely resembled the assemblages found in Indian and South Atlantic samples (Fig. 2A). At higher resolution, barcodes (Fig. 2, B and C) and metagenomic reads (*16*) shared between the Agulhas rings and the Indian or South Atlantic samples showed that the young ring was genetically distinct from both Indian and South Atlantic samples,

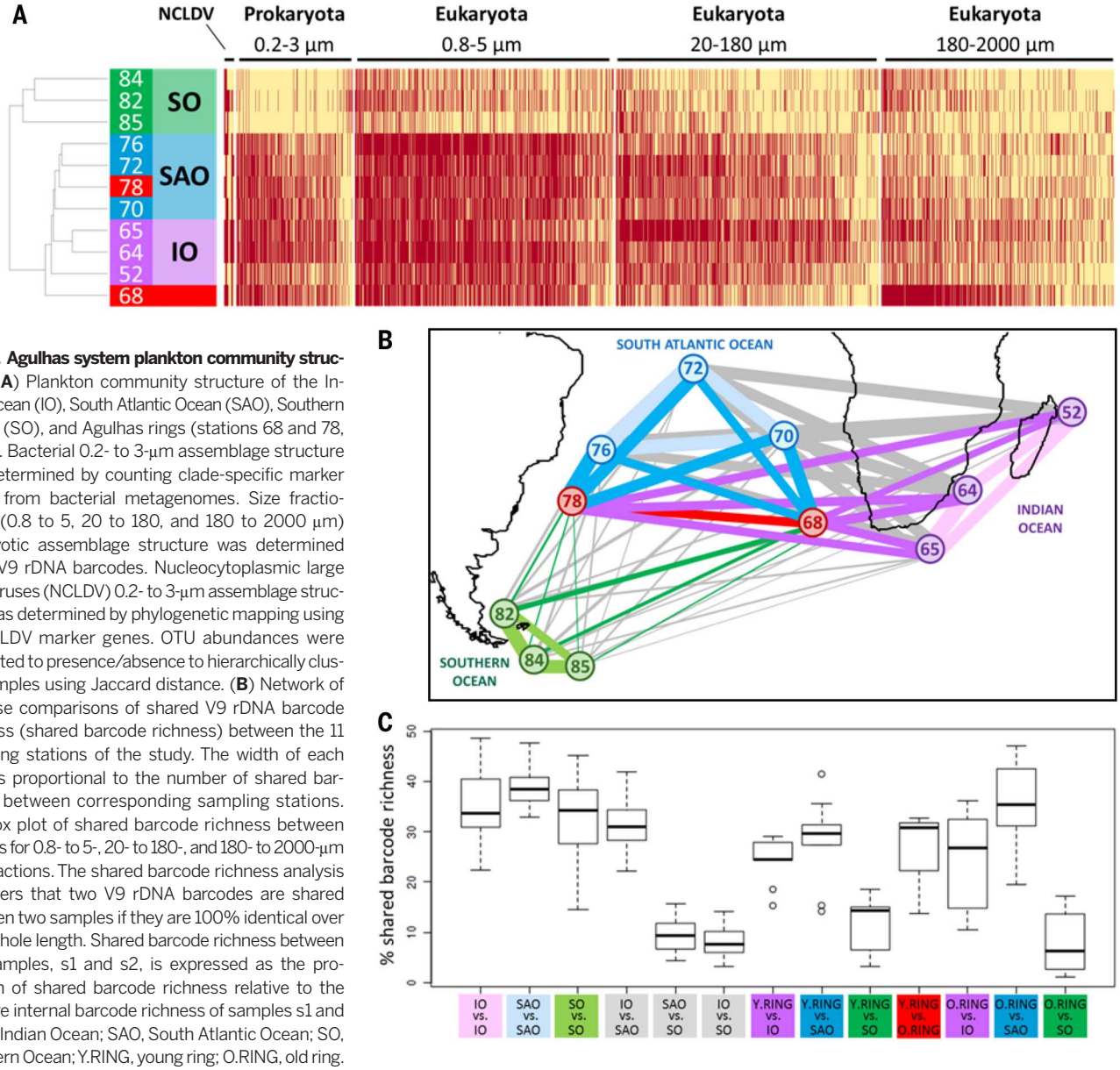


Fig. 2. Agulhas system plankton community structure. (**A**) Plankton community structure of the Indian Ocean (IO), South Atlantic Ocean (SAO), Southern Ocean (SO), and Agulhas rings (stations 68 and 78, in red). Bacterial 0.2- to 3-µm assemblage structure was determined by counting clade-specific marker genes from bacterial metagenomes. Size fractionated (0.8 to 5, 20 to 180, and 180 to 2000 µm) eukaryotic assemblage structure was determined using V9 rDNA barcodes. Nucleocytoplasmic large DNA viruses (NCLDV) 0.2- to 3-µm assemblage structure was determined by phylogenetic mapping using 16 NCLDV marker genes. OTU abundances were converted to presence/absence to hierarchically cluster samples using Jaccard distance. (**B**) Network of pairwise comparisons of shared V9 rDNA barcode richness (shared barcode richness) between the 11 sampling stations of the study. The width of each edge is proportional to the number of shared barcodes between corresponding sampling stations. (**C**) Box plot of shared barcode richness between stations for 0.8- to 5-, 20- to 180-, and 180- to 2000-µm size fractions. The shared barcode richness analysis considers that two V9 rDNA barcodes are shared between two samples if they are 100% identical over their whole length. Shared barcode richness between two samples, s1 and s2, is expressed as the proportion of shared barcode richness relative to the average internal barcode richness of samples s1 and s2. IO, Indian Ocean; SAO, South Atlantic Ocean; SO, Southern Ocean; Y.RING, young ring; O.RING, old ring.

whereas the old ring was similar to its surrounding South Atlantic samples (Tukey post hoc, 0.95 confidence). Light microscopy analyses revealed some plankton groups specific to the young Agulhas ring, such as *Pseudo-nitzschia* spp., which represented 20% of the phytoplankton counts but less than 10% in all other stations (*12*). Other potentially circumstantial plankton characteristic of the young Agulhas ring included the tintinnid *Dictyocysta pacifica* (*12*), the diatom *Corethron pennatum* (*12*), and the dinoflagellate *Tripos limulus* (*12*). A tiny (less than 15 μm long) pennate diatom from the genus *Nanoneis,* which we saw only in the young Agulhas ring and Indian Ocean stations around the African coasts (*28*), was an example of the Indo-Atlantic plankton diversity filtering observed at rDNA barcode level and corroborated by microscopy. OTU clustered barcodes revealed a variety of young Agulhas ring sub-OTU richness patterns compared with source and destination oceans (Fig. 5A). Among Copepoda, *Gaetanus variabilis* and *Corycaeus speciosus* were the more cosmopolitan species (Fig. 5B), whereas *Bradya* species found in the young ring were mainly similar to those from the Indian Ocean. *Acartia negligens* and *Neocalanus robustior* displayed high levels of barcode richness specific to each side of the Agulhas choke point. Bacillariophyceae were heavily filtered from Indian to South Atlantic Oceans (Fig. 5C), and most OTUs (17 out of 20) were absent in the young ring, suggesting that diversity filtering could take place earlier in the ring's 9-month history. Consistent with the observed particularities of the plankton in the young ring, continuous underway optical measurements showed that the ring core photosynthetic community differed from surrounding waters (*29*–*31*). Intermediate size cells, and relatively low content of photoprotective pigments, reflected low growth irradiance and suggested a transitional physiological state. Thus, the plankton community in the young Agulhas ring had diverged from plankton communities typical of its original Indian waters but, even 9 months after formation, had not converged with its surrounding South Atlantic waters.

Deep mixing in Agulhas rings promotes plankton bloom

The upper water column of the young ring showed a high nitrite concentration (>0.5 mmol m^{-3}) (Fig. 4D) (*32*). This observation, along with its particularly deep mixed layer (>250 m), suggested that as Agulhas rings proceed westward in the Cape Basin, vigorous deep mixing of their weakly stratified waters may have entrained nitrate and stimulated phytoplankton blooms. Typically, fresh organic material would then either be exported as sinking particles or locally recycled, sustaining heterotrophic production of ammonium that would, in turn, be consumed by photoautotrophs in the euphotic layer but nitrified below. The resulting nitrite, eventually oxidized to nitrate, might remain evident at subsurface as observed in the nitrite anomaly of the young ring detected here. This hypothesis was supported by numerical simulations of the Massachusetts Institute of Technology General Circulation Model (*33*), which resolved Agulhas rings, their phytoplankton populations, and associated nutrient cycling (Fig. 6A). We tracked 12 Agulhas rings in the ocean model and characterized their near-surface biogeochemical cycles (Fig. 6B) (*34*). As the rings moved westward, storms enhanced surface heat loss, stimulating convection and the entrainment of nitrate. In the model simulations, proliferation of phytoplankton generated subsurface nitrite, which persisted because phytoplankton were light-limited at depth and because nitrification was suppressed by light at the surface (*35*). The associated blooms were dominated by large opportunistic phytoplankton and nitrate-metabolizing *Synechococcus* spp. analogs, whereas populations of *Prochlorococcus* spp. analogs dominated the quiescent periods (*34*). Each of the 12 simulated Agulhas rings exhibited this pattern in response to surface forcing by weather systems, and all rings maintained a persistent

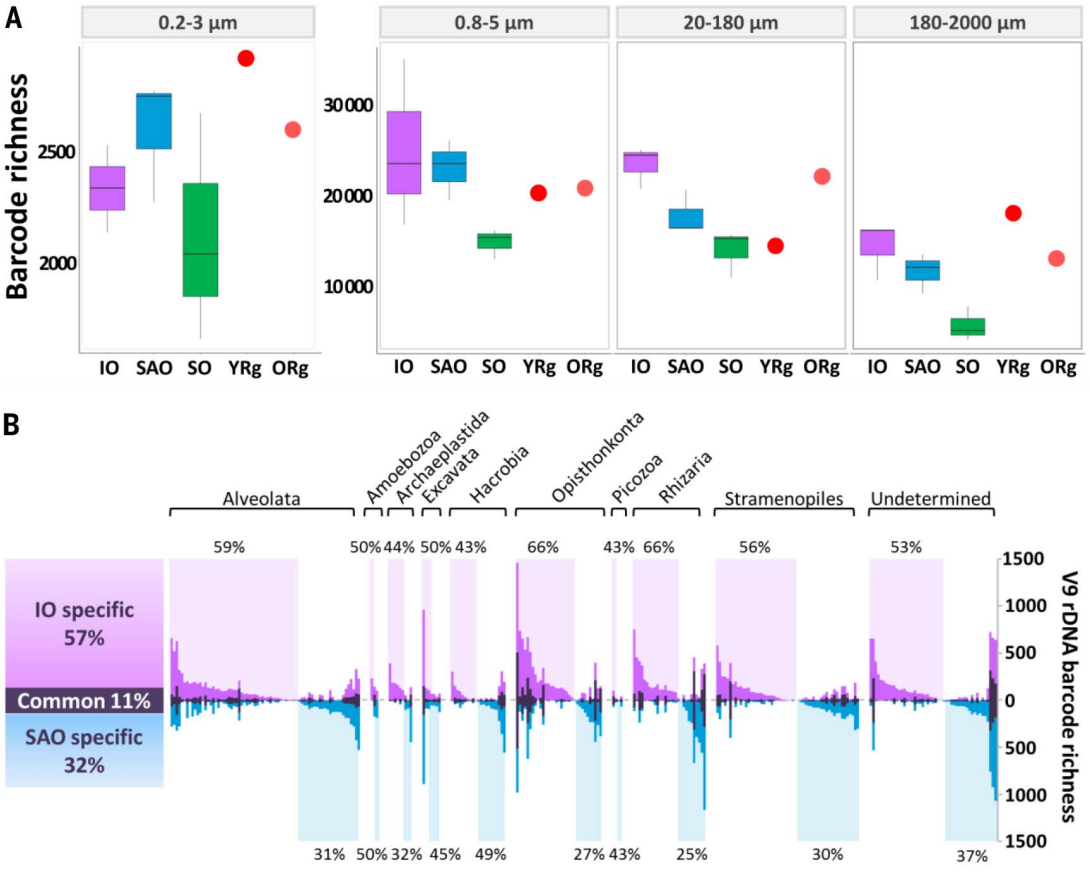


Fig. 3. Diversity of plankton populations specific to Indian and Atlantic Oceans. (**A**) Box plot of 16S (0.2 to 3 μm) and V9 rDNA barcodes richness (0.8- to 5-, 20- to 180-, and 180- to 2000-μm size fractions). Each box represents three sampling stations combined into Indian, South Atlantic, and Southern Ocean. Single Agulhas ring stations are represented as red (young ring) and orange (old ring) crosses. (**B**) Plankton sub-OTU richness filtering across the Agulhas choke point. Each vertical bar represents a single eukaryotic plankton OTU, each of which contains >10 distinct V9 rDNA barcodes (*14*). For each OTU are represented the number of distinct barcodes (sub-OTU richness) found exclusively in the South Atlantic Ocean (blue), exclusively in the Indian Ocean (pink), and in both South Atlantic Ocean and Indian Ocean (gray). OTUs are grouped by taxonomic annotation (indicated above the bar plot). For each taxonomic group, the percentage of OTUs with higher sub-OTU richness in the Indian Ocean (shaded in pink) or in the South Atlantic Ocean (shaded in blue) is indicated, respectively, at the top and bottom of the bar plot. A total of 1307 OTUs are presented, representing 98% of total V9 rDNA barcode abundance.

subsurface nitrite maximum in the region, as observed in TARA_068 and in other biogeochemical surveys (*36*).

The nitrite peak observed at TARA_068 in the young Agulhas ring was associated with a differential representation of nitrogen metabolism genes between the ring and the surrounding South Atlantic and Indian Oceans metagenomes derived from 0.2- to 3-μm size fractions (Fig. 7) (*37*). Agulhas ring overrepresented KEGG (Kyoto Encyclopedia of Genes and Genomes) orthologs (KOs) were involved in both nitrification and denitrification, likely representing the overlap between plankton assemblages involved in the conversion of nitrate to nitrite on the one hand and in denitrification of the accumulating nitrite on the other. Distinct KOs involved in successive denitrification steps were found to be encoded by similar plankton taxa. For instance, KO10945 and KO10946 (involved in ammonium nitrification) and KO00368 (subsequently involved in nitrite to nitrous oxide denitrification) appeared mostly encoded by Nitrosopumilaceae archaea. KO00264 and KO01674 (involved in ammonium assimilation) were mostly assigned to eukaryotic Mamiellales, whereas the opposite KO00367 and KO00366 (involved in dissimilatory nitrite reduction to ammonium), followed by KO01725 (involved in ammonium assimilation), were encoded by picocyanobacteria. In the specific case of the picocyanobacteria, metagenomic reads corresponding to *nirA* genes showed that the observed young Agulhas ring KO00366 (dissimilatory nitrite reduction) enrichment was mainly due to the overrepresentation of genes from *Prochlorococcus* (Fig. 8B). This enrichment was found to be associated with a concomitant shift in population structure from *Prochlorococcus* high-light II ecotypes (HLII, mostly lacking *nirA* genes) to codominance of high-light I (HLI) and low-light I (LLI) ecotypes. Indeed, among the several *Prochlorococcus* and *Synechococcus* ecotypes identified based on their genetic diversity and physiology (*38*, *39*), neutral marker (*petB*) (Fig. 8A) recruitments showed that dominant clades in the Indian Ocean upper mixed layer were *Prochlorococcus* HLII and *Synechococcus* clade II, as expected given the known (sub)tropical preference of these groups (*40*). Both clades nearly completely disappeared (less than 5%) in the mixed cold waters of the young ring and only began to increase again when the surface water warmed up along the South Atlantic Ocean transect. Conversely, young ring water was characterized by a large proportion of *Prochlorococcus* HLI and LLI and *Synechococcus* clade IV, two clades typical of temperate waters. Besides temperature, the *Prochlorococcus* community shift from HLII to HLI + LLI observed in the young ring was likely also driven by the nitrite anomaly. Indeed, whereas most *Synechococcus* strains isolated so far are able to

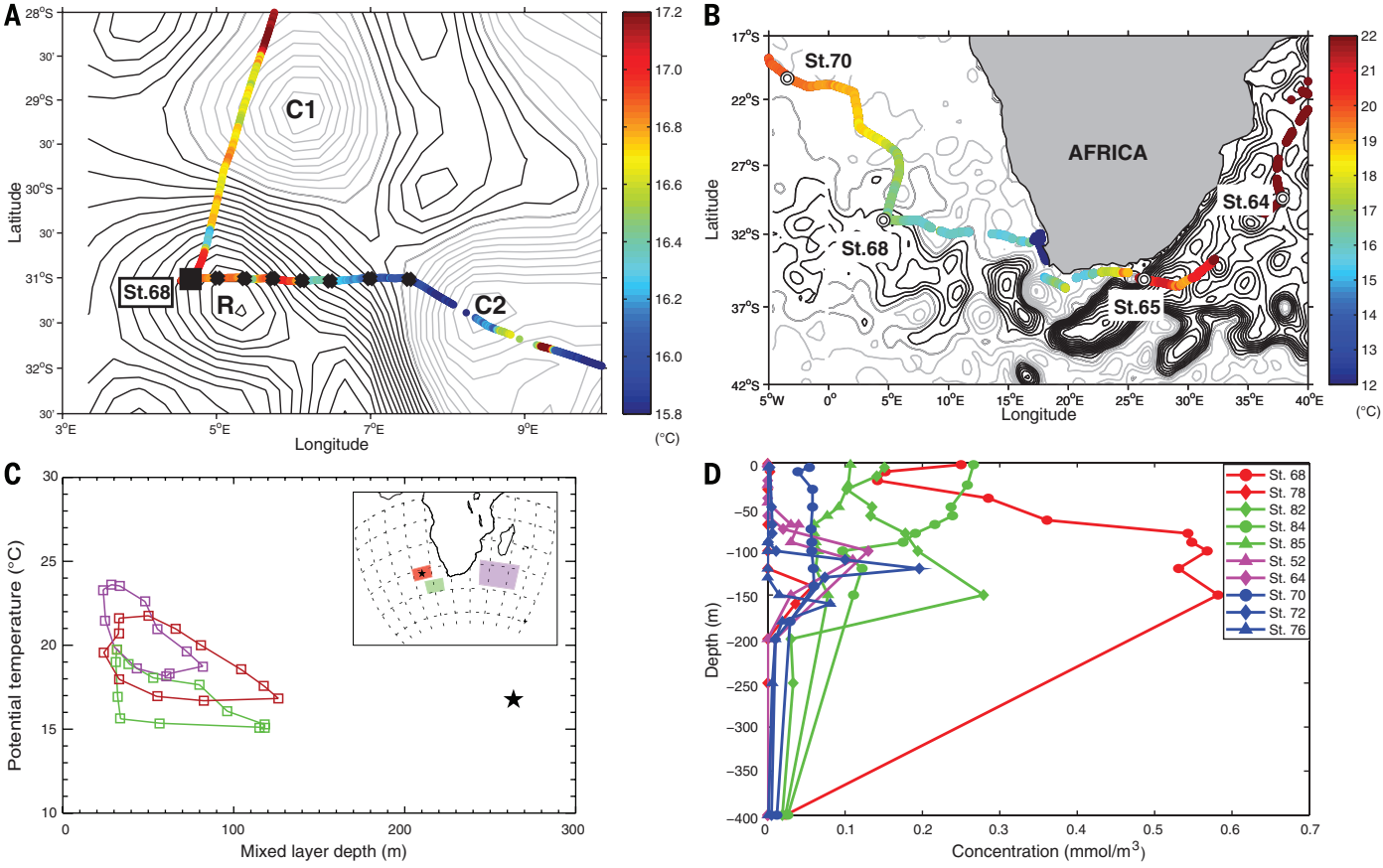


Fig. 4. Properties of the young Agulhas ring (TARA_068). (**A**) Daily sea surface height around young Agulhas ring station TARA_068 [absolute dynamic topography (ADT) from www.aviso.altimetry.fr]. R, C1, and C2, respectively, denote the centers of the Agulhas ring and two cyclonic eddies. The contour interval is 0.02 dyn/m. The ADT values are for 13 September 2010. Light gray isolines, ADT < 0.46 dyn/m. The crosses indicate the CTD stations, and the square symbol indicates the position of the biological station TARA_068. The biological station coincides with the westernmost CTD station. ADT is affected by interpolation errors, which is why CTD casts were performed at sea so as to have a fine-scale description of the feature before defining the position of the biological station (*23*). Superimposed are the continuous underway temperatures (°C) from the on-board thermosalinograph. (**B**) Same as (A) but at the regional scale. Round symbols correspond to biological sampling stations. The contour interval is 0.1 dyn/m. (**C**) Seasonal distribution of the median values of the mixed layer depths and temperatures at 10 m (from ARGO) provided by the IFREMER/LOS Mixed Layer Depth Climatology L2 database (www.ifremer.fr/cerweb/deboyer/mld) updated to 27 July 2011. The mixed layer is defined using a temperature criterion. The star symbol represents the young ring station TARA_068. (Inset) Geographic position of the areas used to select the mixed layer and temperature data. The mixed layer depth measured at TARA_068 is outside the 90th percentile of the distribution of mixed layer depths for the same month for both the subtropical (red and magenta) regions. The temperature matches the median for the same month and region of sampling. (**D**) Nitrite (NO_2) concentrations from CTD casts at different sampling sites (expressed in mmol/m^3).

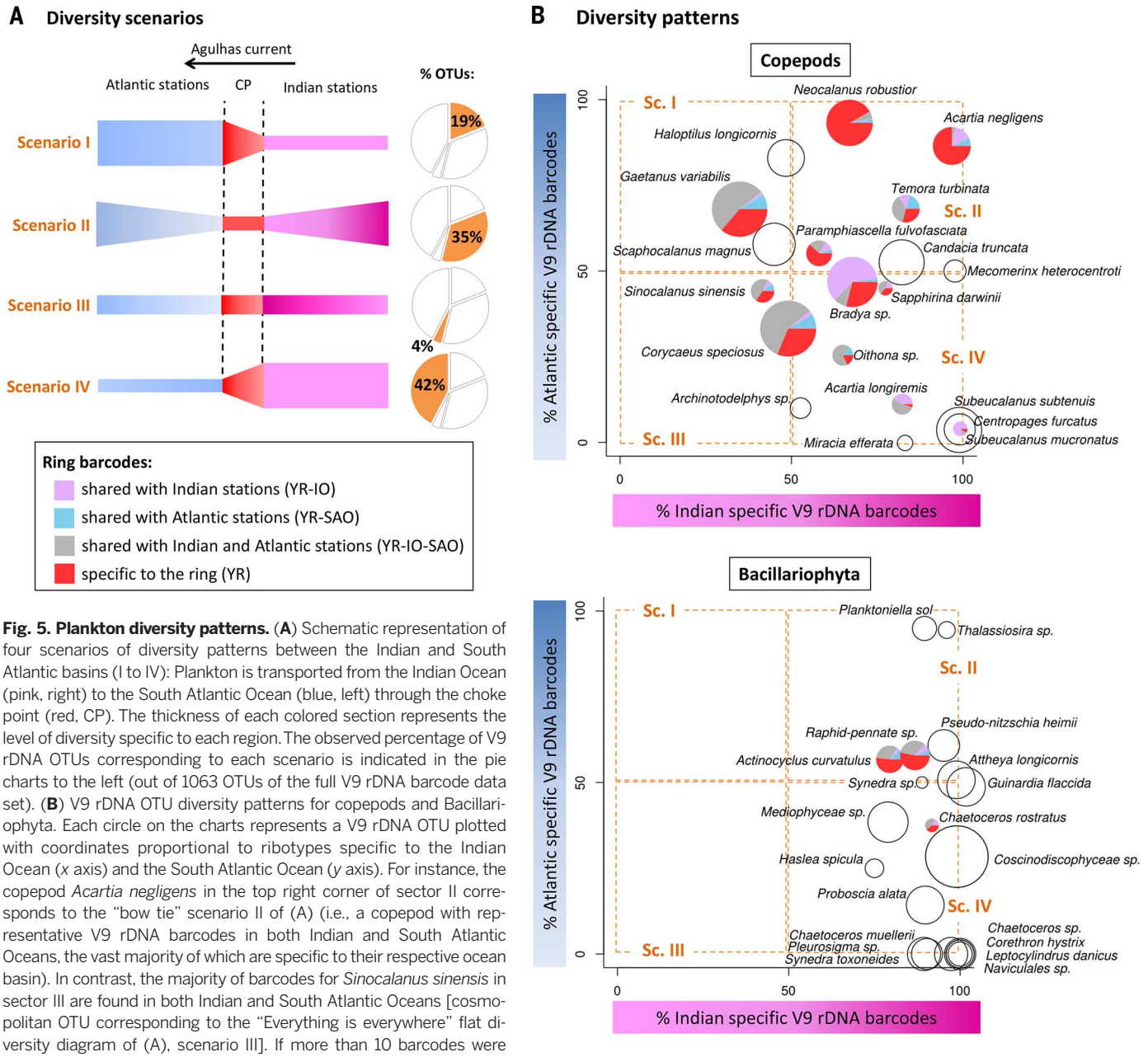


Fig. 5. Plankton diversity patterns. (**A**) Schematic representation of four scenarios of diversity patterns between the Indian and South Atlantic basins (I to IV): Plankton is transported from the Indian Ocean (pink, right) to the South Atlantic Ocean (blue, left) through the choke point (red, CP). The thickness of each colored section represents the level of diversity specific to each region. The observed percentage of V9 rDNA OTUs corresponding to each scenario is indicated in the pie charts to the left (out of 1063 OTUs of the full V9 rDNA barcode data set). (**B**) V9 rDNA OTU diversity patterns for copepods and Bacillariophyta. Each circle on the charts represents a V9 rDNA OTU plotted with coordinates proportional to ribotypes specific to the Indian Ocean (*x* axis) and the South Atlantic Ocean (*y* axis). For instance, the copepod *Acartia negligens* in the top right corner of sector II corresponds to the "bow tie" scenario II of (A) (i.e., a copepod with representative V9 rDNA barcodes in both Indian and South Atlantic Oceans, the vast majority of which are specific to their respective ocean basin). In contrast, the majority of barcodes for *Sinocalanus sinensis* in sector III are found in both Indian and South Atlantic Oceans [cosmopolitan OTU corresponding to the "Everything is everywhere" flat diversity diagram of (A), scenario III]. If more than 10 barcodes were found in the young Agulhas ring (TARA_068), their distribution is indicated in a pie chart (colors are coded in the legend inset); otherwise, the OTU is represented by an empty circle. Circle sizes are proportional to the number of considered barcodes for each OTU. The Bacillariophyta OTU defined as *Raphid pennate* sp. likely corresponds to the *Pseudo-nitzschia* cells observed by light microscopy.

use nitrate, nitrite, and ammonium, only the *Prochlorococcus* LLI and IV and some populations of HL clades, having acquired the *nirA* gene by lateral gene transfer, are able to assimilate nitrite. In the young ring, overrepresentation of cyanobacterial orthologs involved in nitrite reduction could thus have resulted from environmental pressure selecting LLI (87% of the *nirA* recruitments) and HL populations (13%) that possessed this ability. Because the capacity to assimilate nitrite in this latter ecotype reflects the availability of this nutrient in the environment (*41*), these in situ observations of picocyanobacteria indicated that the nitrogen cycle disturbance occurring in the young ring exerts community-wide selective pressure on Agulhas ring plankton.

Discussion

We found that whether or not the Agulhas choke point is considered a barrier to plankton dispersal depends on the taxonomic resolution at which the analysis is performed. At coarse taxonomic resolution, our observations of Indo-Atlantic continuous plankton structure—from viruses to fish larvae—suggested unlimited dispersal, consistent with previous reports (*5*, *42*). However, at finer resolution, our genetic data revealed that the Agulhas choke point strongly affects patterns of plankton genetic diversity. As anticipated in (*5*), the diversity filtering by Agulhas rings likely escaped detection using fossil records because of the limited taxonomic resolution afforded by fossil diatom morphology (*42*). The community-wide evidence presented here confirms observations on individual living species (*43*, *44*), suggesting that dispersal filters mitigate the panmictic ocean hypothesis for plankton above 20 μm.

The lower diversity we observed in the South Atlantic Ocean for micro- and mesoplankton (>20 μm) may be due to local abiotic/biotic pressure or to limitations in dispersal (*33*, *45*). Biogeography emerging from a model with only neutral drift (*46*) predicts

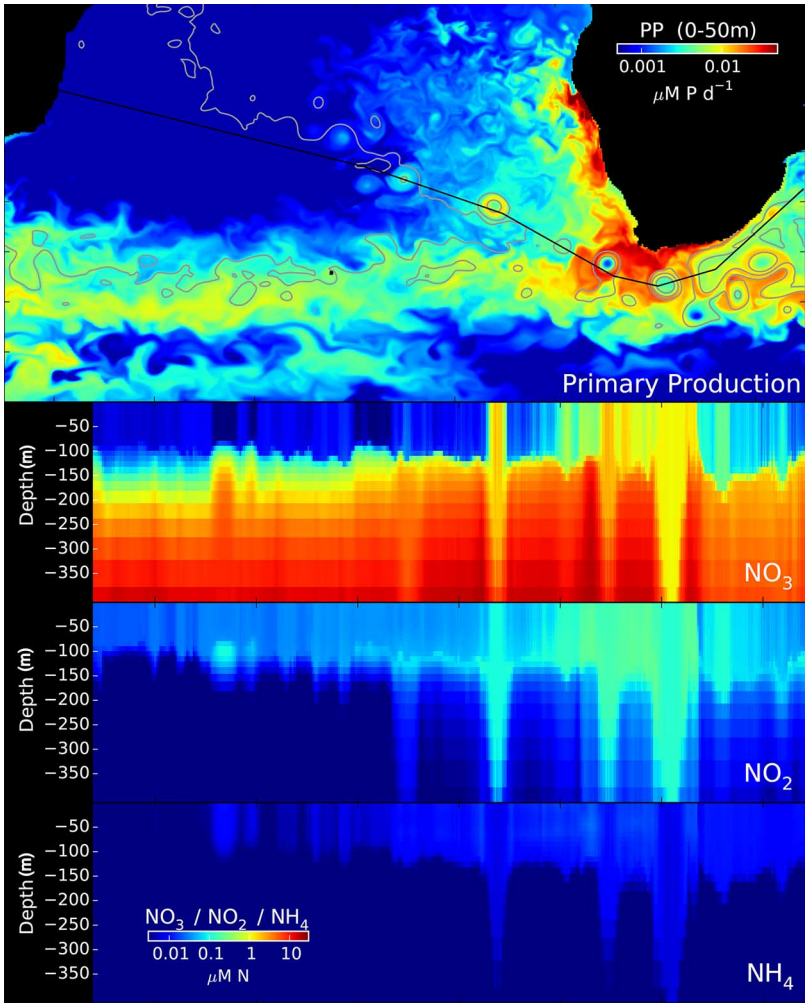


Fig. 6. Modeled nitrogen stocks along Agulhas ring track. (**Top**) Simulated primary production (PP) in the Agulhas system using the MIT-GCM model. The solid black line shows the average northwesterly path of 12 distinct virtual Agulhas rings tracked over the course of the simulation. Color scale for PP is given in the top right inset, with warmer colors indicating higher PP. (**Bottom**) Modeled profiles of NO_3, NO_2, and NH_4 along the Agulhas ring average track (*x* axis) presented in (A). The *y* axis is the depth (in meters) in the water column. The color scale is given in the bottom left inset, with warmer colors indicating higher concentrations of nitrogen compounds.

basin-to-basin genetic differences that are qualitatively consistent with our data. However, the increased proportion of *Prochlorococcus* HL populations carrying the *nirA* gene in the young Agulhas ring indicates that selection is at work in Agulhas rings. Based on our analysis of two Agulhas rings, we propose that environmental disturbances in Agulhas rings reshape their plankton diversity as they travel from the Indian Ocean to the South Atlantic Ocean. Such selective pressure may contribute to the South Atlantic Ocean plankton diversity shift relative to its upstream Indo-Pacific basin. Thus, environmental selection applied at a choke point in ocean circulation may constitute a barrier to dispersal (*47*, *48*). Furthermore, we show that taxonomic groups were not equally affected by the ring transport, both within and between phyla, with a noticeable effect of organism size. The differential effects due to organism size highlight the difficulty in generalizing ecological and evolutionary rules from limited sampling of species or functional types.

Considering the sensitivity of Agulhas leakage to climate change (*1*, *49*), better understanding of the plankton dynamics in Agulhas rings will be required if we are to understand and predict ecosystem resilience at the planetary scale. Considering the breadth of changes already observed in the 9-month-old Agulhas ring, it would be interesting to acquire samples from specific Agulhas rings tracked from early formation to dissipation. Finally, our data suggest that the abundance of Indian Ocean species in South Atlantic Ocean sedimentary records, used as proxies of Agulhas leakage intensity (*4*), may actually also depend on the physical and biological characteristics of the Agulhas rings.

Materials and methods

Sampling

The *Tara* Oceans sampling protocols schematized in Karsenti *et al.* (*9*) are described in Pesant *et al.* (*50*); specific methods for 0.8- to 5-, 20- to 180-, and 180- to 2000-μm size fractions in de Vargas *et al.* (*17*); for 0.2- to 3-μm size fractions in Sunagawa *et al.* (*51*); and for <0.2-μm size fraction in Brum *et al.* (*52*). Due to their fragility, 1.6-μm glass fiber filters initially used for prokaryote sampling were replaced by more resistant 3-μm polycarbonate filters from station TARA_066 onward. In the present text, both 0.2- to 1.6-μm and 0.2- to 3-μm prokaryote size fractions are simply referred to as 0.2 to 3 μm.

Data acquisition

A range of analytical methods covering different levels of taxonomic resolution (pigments, flow cytometry, optical microscopy, marker gene barcodes, and metagenomics) were used to describe the planktonic composition at each sampled station. Viruses from the <0.2 μm size fraction were studied by epifluorescence microscopy, by quantitative transmission electron microscopy, and by sequencing DNA as described in Brum *et al.* (*52*). Flow cytometry was used to discriminate high-DNA-content bacteria (HNA), low-DNA-content bacteria (LNA), *Prochlorococcus* and *Synechococcus* picocyanobacteria, and two different groups (based on their size) of photosynthetic picoeukaryotes, as described previously (*53*). Pigment concentrations measured by high-performance liquid chromatography (HPLC) were used to estimate the dominant classes of phytoplankton using the CHEMTAX procedure (*54*). Tintinnids, diatoms, and dinoflagellates were identified and counted by light microscopy from the 20- to 180-μm lugol or formaldehyde fixed-size fraction. Zooplankton enumeration was performed on formol fixed samples using the ZOOSCAN semi-automated classification of digital images (*55*). Sequencing, clustering, and annotation of 18*S*-V9 rDNA barcodes are described in de Vargas *et al.* (*17*). Metagenome sequencing, assembly, and annotation are described in Sunagawa *et al.* (*51*). NCLDV taxonomic assignations in the 0.2- to 3-μm samples were carried out using 18 lineage-specific markers as described in Hingamp *et al.* (*56*). Virome sequencing and annotation are described in Brum *et al.* (*52*). Samples and their associated contextual data are described at PANGAEA (*57*–*59*).

Data analysis

Origin of sampled Agulhas rings

Using visual and automated approaches, the origins of the TARA_068 and TARA_078 stations were traced back from the daily altimetric data (Fig. 1) (*21*). The automated approach used either the Lagrangian tracing of numerical particles initialized in the center of a given structure and transported by the geostrophic velocity field calculated from sea surface height gradients, or the connection in space and time of adjacent extreme values in sea level anomaly maps.

V9 rDNA barcodes

To normalize for differences in sequencing effort, V9 rDNA barcode libraries were resampled 50 times for the number of reads corresponding to the smallest library in each size fraction: 0.8 to 5 μm, 776,358 reads; 20 to 180 μm, 1,170,592 reads; and 180 to 2000 μm, 767,940 reads. V9 rDNA barcode counts were then converted to the average number

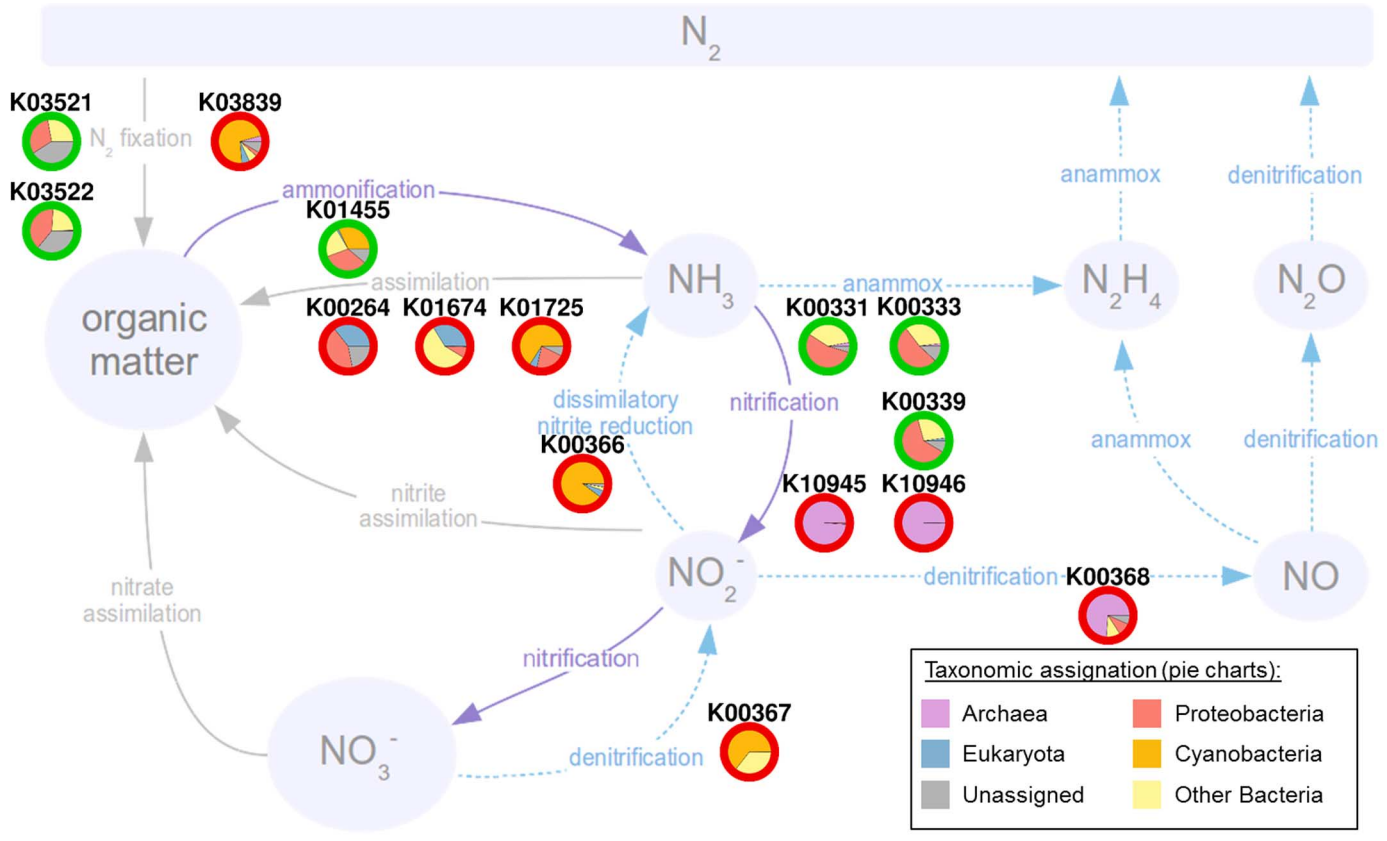


Fig. 7. Nitrite anomaly in the young Agulhas ring is accompanied by shifts in nitrogen pathway–related genes. Metagenomic over- and underrepresented nitrogen pathway genes in young Agulhas ring. Over- (red circles) and under- (green circles) represented metagenome functional annotations (KEGG Orthologs, KO#) involved in the nitrogen pathway in the young ring compared to Indian and South Atlantic Oceans reference stations, at surface and deep chlorophyll maximum depth. Pie charts inside circles represent the taxonomic distribution for each ortholog.

of times seen in the 50 resampling events, and barcodes with less than 10 reads were removed as potential sequencing artifacts. We used downsampled barcode richness (number of distinct V9 rDNA barcodes) as a diversity descriptor because using V9 rDNA barcode abundances to compare plankton assemblages would likely be biased due to (i) technical limitations described in de Vargas *et al.* (*17*) and (ii) seasonality effects induced by the timing of samplings (table S1). Barcode richness was well correlated with Shannon and Simpson indexes (0.94 and 0.78, respectively). The shared barcode richness between each pair of samples (*14*) was estimated by counting, for the three larger size fractions (0.8 to 5, 20 to 180, and 180 to 2000 μm), the proportion of V9 rDNA barcodes 100% identical over their whole length. V9 rDNA barcodes were clustered into OTUs by swarm clustering as described by de Vargas *et al.* (*17*). The sub-OTU richness comparison between two samples s1 and s2 (*14*) produces three values: the number of V9 rDNA barcodes in common, the number of V9 rDNA barcodes unique to s1, and the number of V9 rDNA barcodes unique to s2. These numbers can be represented directly as bar graphs (Fig. 3B) or as dot plots of specific V9 rDNA barcode richness (Fig. 5).

Metagenomic analysis

Similarity was estimated using whole shotgun metagenomes for all four available size fractions (0.2 to 3, 0.8 to 5, 20 to 180, and 180 to 2000 μm). Because pairwise comparisons of all raw metagenome reads are intractable given the present data volume, we used a heuristic in which two metagenomic 100–base pair (bp) reads were considered similar if at least two nonoverlapping 33-bp subsequences were strictly identical (Compareads method) (*60*). For prokaryotic fractions (0.2 to 3 μm), taxonomic abundance was estimated using the number of $16S_{mi}$tags (*51*). The functional annotation, taxonomic assignation, and gene abundance estimation of the panoceanic Ocean Microbial Reference Gene Catalog (OM-RGC) (243 samples, including all those analyzed here) generated from *Tara* Oceans 0.2- to 3-μm metagenomic reads are described in Sunagawa *et al.* (*51*). Gene abundances were computed for the set of genes annotated to the nitrogen metabolism KO (*61*) group by counting the number of reads from each sample that mapped to each KO-associated gene. Abundances were normalized as reads per kilobase per million mapped reads (RPKM). Gene abundances were then aggregated (summed) for each KO group. To compare abundances between the young ring (TARA_068) and other stations, a *t* test was used. KOs with a *P* value <0.05 and a total abundance (over all stations) >10 were considered as significant (*37*). *Prochlorococcus* and *Synechococcus* community composition was analyzed in the 0.2- to 3-μm size fraction at the clade level by recruiting reads targeting the high-resolution marker gene *petB*, coding for cytochrome b_6 (*62*). The *petB* reads were first extracted from metagenomes using Basic Local Alignment Search Tool (BLASTx+) against the *petB* sequences of *Synechococcus* sp. WH8102 and *Prochlorococcus marinus* MED4. These reads were subsequently aligned against a reference data set of 270 *petB* sequences using BLASTn (with parameters set at -G 8 -E 6 -r 5 -q -4 -W 8 -e 1 -F "m L" -U T). *petB* reads exhibiting >80% identity over >90% of sequence length were then taxonomically assigned to the clade of the best BLAST hit. Read counts per clade were normalized based on the sequencing effort for each metagenomic sample. A similar approach was used with *nirA* (KO 00366) and *narB* genes (KO 00367), which were highlighted in the nitrogen-related KO analysis (Fig. 7). Phylogenetic assignment was realized at the highest possible taxonomic level using a reference data set constituted of sequences retrieved from Cyanorak v2 (www.sb-roscoff.fr/cyanorak/) and Global Ocean Sampling (*41*, *63*) databases.

Nitrogen cycle modeling

Numerical simulations of global ocean circulation were based on the Massachusetts Institute of Technology General Circulation Model (MIT-GCM) (*64*), incorporating biogeochemical and ecological components (*65*, *66*). It resolved mesoscale

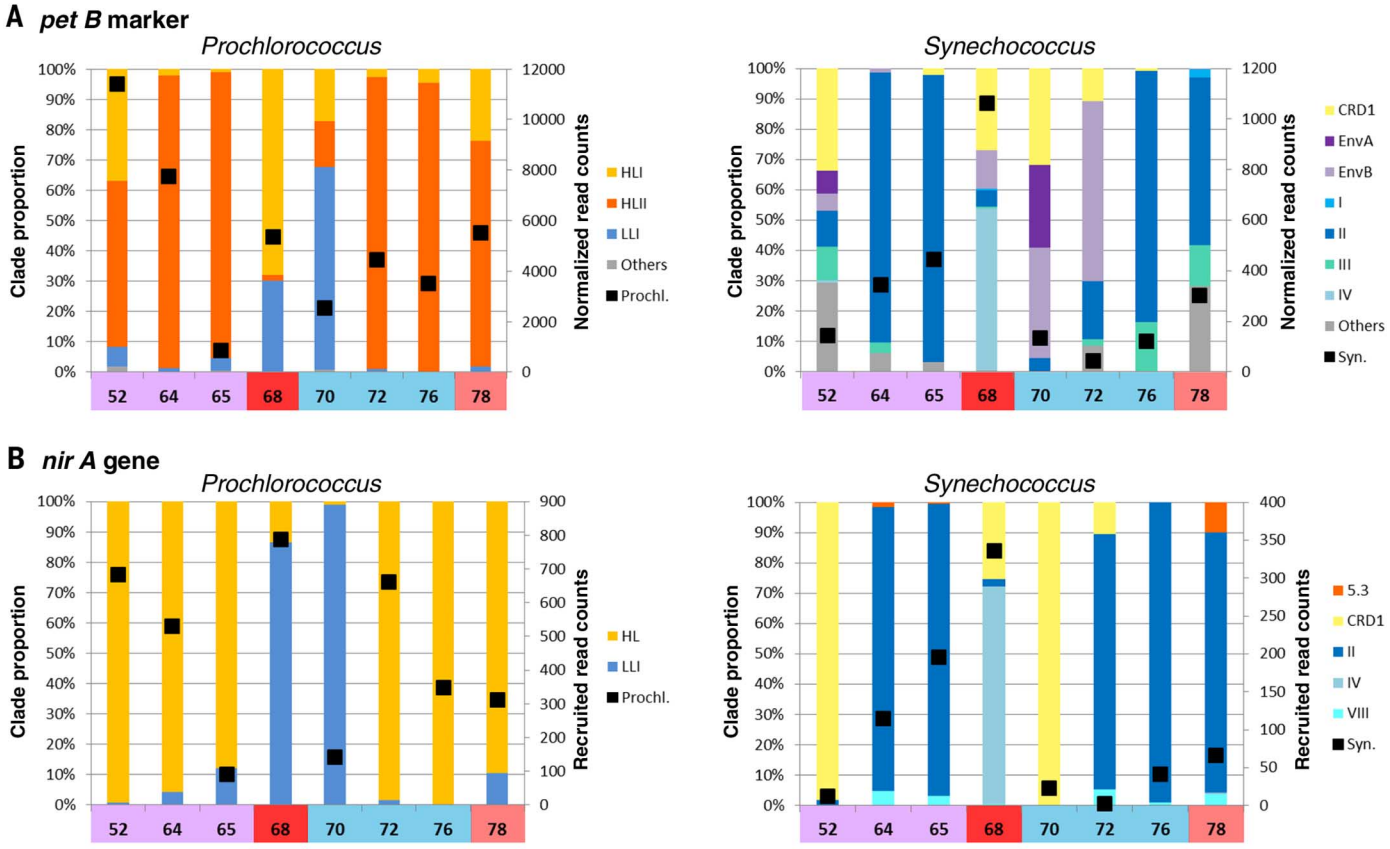


Fig. 8. Picocyanobacterial clade shift in the young Agulhas ring. (**A**) Relative abundance of *Prochlorococcus* and *Synechococcus* clades, estimated by *petB* read recruitments from 0.2- to 3-µm metagenomes. Solid squares correspond to read counts normalized based on the sequencing effort (right axis). (**B**) Relative abundance of *nirA* gene from *Prochlorococcus* and *Synechococcus* clades estimated by number of reads recruited from 0.2- to 3-µm metagenomes. The bar colors correspond to cyanobacterial clades indicated in the inset legends for each panel. Solid squares correspond to the number of reads recruited (right axis). Data are shown for stations TARA_052 to TARA_078 only, because too few cyanobacteria were found in Southern Ocean stations TARA_082, TARA_084, and TARA_085.

features in the tropics and was eddy-permitting in subpolar regions. The physical configurations were integrated from 1992 to 1999 and constrained to be consistent with observed hydrography and altimetry (*67*). Three inorganic fixed nitrogen pools were resolved—nitrate, nitrite, and ammonium—as well as particulate and dissolved detrital organic nitrogen. Phytoplankton types were able to use some or all of the fixed nitrogen pools. Aerobic respiration and remineralization by heterotrophic microbes was parameterized as a simple sequence of transformations from detrital organic nitrogen, to ammonium, then nitrification to nitrite and nitrate. In accordance with empirical evidence (*35*), nitrification was assumed to be inhibited by light. Nitrification is described in the model by simple first-order kinetics, with rates tuned to qualitatively capture the patterns of nitrogen species in the Atlantic (*66*).

Continuous spectral analysis

A continuous flow-through system equipped with a high-spectral-resolution spectrophotometer (AC-S, WET Labs, Inc.) was used for data collection during the *Tara* Oceans expedition, as described previously (*68*). Phytoplankton pigment concentrations, estimates of phytoplankton size γ, total chlorophyll a concentration, and particulate organic carbon (POC) are derived from the absorption and attenuation spectra (*69*) for the 1-km^2-binned *Tara* Oceans data set available at PANGAEA (http://doi.pangaea.de/10.1594/PANGAEA.836318).

REFERENCES AND NOTES

1. A. Biastoch, C. W. Böning, J. R. E. Lutjeharms, Agulhas leakage dynamics affects decadal variability in Atlantic overturning circulation. *Nature* **456**, 489–492http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=19037313&dopt=Abstract (2008). doi: 10.1038/nature07426
2. A. Biastoch, C. W. Böning, F. U. Schwarzkopf, J. R. E. Lutjeharms, Increase in Agulhas leakage due to poleward shift of Southern Hemisphere westerlies. *Nature* **462**, 495–498 http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=19940923&dopt=Abstract (2009). doi: 10.1038/nature08519
3. L. M. Beal *et al*., On the role of the Agulhas system in ocean circulation and climate. *Nature* **472**, 429–436 (2011). doi: 10.1038/nature09983; pmid: 21525925
4. F. J. C. Peeters *et al*., Vigorous exchange between the Indian and Atlantic oceans at the end of the past five glacial periods. *Nature* **430**, 661–665 (2004). doi: 10.1038/nature02785; pmid: 15295596
5. P. Cermeño, P. G. Falkowski, Controls on diatom biogeography in the ocean. *Science* **325**, 1539–1541 (2009). pmid: 19762642
6. A. L. Gordon, Oceanography: The brawniest retroflection. *Nature* **421**, 904–905 (2003). doi: 10.1038/421904a; pmid: 12606984
7. H. M. van Aken *et al*., Observations of a young Agulhas ring, Astrid, during MARE in March 20. *Deep Sea Res. Part II Top. Stud. Oceanogr.* **50**, 167–195 (2003). doi: 10.1016/S0967-0645(02)00383-1
8. J. R. Bernhardt, H. M. Leslie, Resilience to climate change in coastal marine ecosystems. *Annu. Rev. Mar. Sci.* **5**, 371–392 (2013). doi: 10.1146/annurev-marine-121211-172411; pmid: 22809195
9. E. Karsenti *et al*., A holistic approach to marine eco-systems biology. *PLOS Biol.* **9**, e1001177 (2011). doi: 10.1371/journal.pbio.1001177; pmid: 22028628
10. Companion Web site, tables W2 and W3; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#TablesW
11. A. H. Orsi, T. Whitworth III, W. D. Nowlin Jr., On the meridional extent and fronts of the Antarctic Circumpolar Current. *Deep-Sea Res.* **42**, 641–673 (1995). doi: 10.1016/0967-0637(95)00021-W
12. Companion Web site, tables W4 to W12; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#TablesW
13. Companion Web site, figure W1; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW1
14. Companion Web site, figure W2; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW2
15. Companion Web site, tables W13 and W14; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#TablesW
16. Companion Web site, figure W3; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW3
17. C. de Vargas *et al*., Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).
18. B. W. Bowen, L. A. Rocha, R. J. Toonen, S. A. KarlToBo Laboratory, The origins of tropical marine biodiversity. *Trends Ecol. Evol.* **28**, 359–366 (2013). doi: 10.1016/j.tree.2013.01.018; pmid: 23453048
19. R. L. Cunha *et al*., Ancient divergence in the trans-oceanic deep-sea shark *Centroscymnus crepidater*. *PLOS ONE* **7**, e49196 (2012). doi: 10.1371/journal.pone.0049196; pmid: 23145122
20. C. Schmid *et al*., Early evolution of an Agulhas Ring. *Deep Sea Res. Part II Top. Stud. Oceanogr.* **50**, 141–166 (2003). doi: 10.1016/S0967-0645(02)00382-X

21. Companion Web site, figure W4; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW4
22. Companion Web site, figure W5; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW5
23. Companion Web site, figure W6; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW6
24. Companion Web site, figure W7; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW7
25. L. Gordon, J. R. Lutjeharms, M. L. Gründlingh, Stratification and circulation at the Agulhas Retroflection. *Deep-Sea Res. A, Oceanogr. Res. Pap.* **34**, 565–599 (1987). doi: 10.1016/0198-0149(87)90006-9
26. V. Faure, M. Arhan, S. Speich, S. Gladyshev, Heat budget of the surface mixed layer south of Africa. *Ocean Dyn.* **61**, 1441–1458 (2011). doi: 10.1007/s10236-011-0444-1
27. Companion Web site, figure W8; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW8
28. Companion Web site, figure W9; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW9
29. Companion Web site, text W1; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#TextW1
30. Companion Web site, figure W10; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW10
31. Companion Web site, figure W11; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW11
32. Companion Web site, figure W12; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW12
33. S. Clayton, S. Dutkiewicz, O. Jahn, M. J. Follows, Dispersal, eddies, and the diversity of marine phytoplankton. *Limnol. Oceanogr. Fluids Environ.* **3**, 182–197 (2013). doi: 10.1215/21573689-2373515
34. Companion Web site, figure W13; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW13
35. R. J. Olson, Differential photoinhibition of marine nitrifying bacteria: A possible mechanism for the formation of the primary nitrite maximum. *J. Mar. Res.* **39**, 227–238 (1981).
36. S. Levitus *et al*., The World Ocean Database. *Data Sci. J.* **12**, WDS229–WDS234 (2013). doi: 10.2481/dsj.WDS-041
37. Companion Web site, table W15; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#TablesW
38. D. J. Scanlan *et al*., Ecological genomics of marine picocyanobacteria. *Microbiol. Mol. Biol. Rev.* **73**, 249–299 (2009). doi: 10.1128/MMBR.00035-08; pmid: 19487728
39. Z. I. Johnson *et al*., Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. *Science* **311**, 1737–1740 (2006). doi: 10.1126/science.1118052; pmid: 16556835
40. K. Zwirglmaier *et al*., Global phylogeography of marine Synechococcus and Prochlorococcus reveals a distinct partitioning of lineages among oceanic biomes. *Environ. Microbiol.* **10**, 147–161 (2008). pmid: 17900271
41. A. C. Martiny, S. Kathuria, P. M. Berube, Widespread metabolic potential for nitrite and nitrate assimilation among *Prochlorococcus* ecotypes. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 10787–10792 (2009). doi: 10.1073/pnas.0902532106; pmid: 19549842
42. C. Hubert *et al*., A constant flux of diverse thermophilic bacteria into the cold Arctic seabed. *Science* **325**, 1541–1544 (2009). doi: 10.1126/science.1174012; pmid: 19762643
43. C. K. C. Churchill, Á. Valdés, D. Ó. Foighil, Afro-Eurasia and the Americas present barriers to gene flow for the cosmopolitan neustonic nudibranch *Glaucus atlanticus*. *Mar. Biol.* **161**, 899–910 (2014). doi: 10.1007/s00227-014-2389-7
44. N. Selje, M. Simon, T. Brinkhoff, A newly discovered *Roseobacter* cluster in temperate and polar oceans. *Nature* **427**, 445–448 (2004). doi: 10.1038/nature02272; pmid: 14749832
45. G. Casteleyn *et al*., Limits to gene flow in a cosmopolitan marine planktonic diatom. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 12952–12957 (2010). doi: 10.1073/pnas.1001380107; pmid: 20615950
46. F. L. Hellweger, E. van Sebille, N. D. Fredrick, Biogeographic patterns in ocean microbes emerge in a neutral agent-based model. *Science* **345**, 1346–1349 (2014). doi: 10.1126/science.1254421; pmid: 25214628
47. D. H. Janzen, Why mountain passes are higher in the tropics. *Am. Nat.* **101**, 233–249 (1967). doi: 10.1086/282487
48. G. Wang, M. E. Dillon, Recent geographic convergence in diurnal and annual temperature cycling flattens global thermal profiles. *Nature Climate Change* **4**, 988–992 (2014). doi: 10.1038/nclimate2378
49. C. Backeberg, P. Penven, M. Rouault, Impact of intensified Indian Ocean winds on mesoscale variability in the Agulhas system. *Nature Clim. Change* **2**, 608–612 (2012). doi: 10.1038/nclimate1587
50. S. Pesant *et al*., Open science resources for the discovery and analysis of *Tara* Oceans data. http://biorxiv.org/content/early/2015/05/08/019117 (2015).
51. S. Sunagawa *et al*., Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
52. J. R. Brum *et al*., Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498 (2015).
53. J. M. Gasol, P. A. Del Giorgio, Using flow cytometry for counting natural planktonic bacteria and understanding the structure of planktonic bacterial communities. *Sci. Mar.* **64**, 197–224 (2000).
54. M. Mackey, D. Mackey, H. Higgins, S. Wright, CHEMTAX - a program for estimating class abundances from chemical markers: Application to HPLC measurements of phytoplankton. *Mar. Ecol. Prog. Ser.* **144**, 265–283 (1996). doi: 10.3354/meps144265
55. G. Gorsky *et al*., Digital zooplankton image analysis using the ZooScan integrated system. *J. Plankton Res.* **32**, 285–303 (2010). doi: 10.1093/plankt/fbp124
56. P. Hingamp *et al*., Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J.* **7**, 1678–1695 (2013). doi: 10.1038/ismej.2013.59; pmid: 23575371
57. Tara Oceans Consortium Coordinators; Tara Oceans Expedition, Participants (2014): Registry of selected samples from the Tara Oceans Expedition (2009-2013). doi: 10.1594/PANGAEA.840721
58. S. Chaffron, L. Guidi, F. D'Ovidio, S. Speich, S. Audic, S. De Monte, D. Iudicone, M. Picheral, S. Pesant; Tara Oceans Consortium Coordinators, Tara Oceans Expedition, Participants (2014): Contextual environmental data of selected samples from the Tara Oceans Expedition (2009-2013). doi: 10.1594/PANGAEA.840718
59. S. Chaffron, F. D'Ovidio, S. Sunagawa, S. G. Acinas, L. P. Coelho, S. De Monte, G. Salazar, S. Pesant; Tara Oceans Consortium Coordinators, Tara Oceans Expedition, Participants (2014): Contextual biodiversity data of selected samples from the Tara Oceans Expedition (2009-2013). doi: 10.1594/PANGAEA.840698
60. N. Maillet, C. Lemaitre, R. Chikhi, D. Lavenier, P. Peterlongo, Compareads: Comparing huge metagenomic experiments. *BMC Bioinformatics* **13** (suppl. 19), S10 (2012). doi: 10.1186/1471-2105-13-S19-S10; pmid: 23282463
61. M. Kanehisa *et al*., KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36** (Database), D480–D484 (2008). doi: 10.1093/nar/gkm882; pmid: 18077471
62. S. Mazard, M. Ostrowski, F. Partensky, D. J. Scanlan, Multi-locus sequence analysis, taxonomic resolution and biogeography of marine Synechococcus. *Environ. Microbiol.* **14**, 372–386 (2012). doi: 10.1111/j.1462-2920.2011.02514.x; pmid: 21651684
63. D. B. Rusch *et al*., The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLOS Biol.* **5**, e77 (2007). doi: 10.1371/journal.pbio.0050077; pmid: 17355176
64. J. Marshall, A. Adcroft, C. Hill, L. Perelman, C. Heisey, A finite-volume, incompressible Navier Stokes model for studies of the ocean on parallel computers. *J. Geophys. Res.* **102** (C3), 5753–5766 (1997). doi: 10.1029/96JC02775
65. M. J. Follows, S. Dutkiewicz, S. Grant, S. W. Chisholm, Emergent biogeography of microbial communities in a model ocean. *Science* **315**, 1843–1846 (2007). doi: 10.1126/science.1138544; pmid: 17395828
66. S. Dutkiewicz, M. J. Follows, J. G. Bragg, Modeling the coupling of ocean ecology and biogeochemistry. *Global Biogeochem. Cycles* **23**, GB4017 (2009). doi: 10.1029/2008GB003405
67. D. Menemenlis *et al*., ECCO2: High resolution global ocean and sea ice data synthesis. *Mercator Ocean Quarterly Newsletter* **31**, 13–21 (2008).
68. A. Chase *et al*., Decomposition of in situ particulate absorption spectra. *Methods in Oceanography.* **7**, 110–124 (2013). doi: 10.1016/j.mio.2014.02.002
69. E. Boss *et al*., The characteristics of particulate absorption, scattering and attenuation coefficients in the surface ocean; Contribution of the Tara Oceans expedition. *Methods in Oceanography.* **7**, 52–62 (2013). doi: 10.1016/j.mio.2013.11.002
70. K. R. Ridgway, J. R. Dunn, J. L. Wilkin, Ocean interpolation by four-dimensional least squares - Application to the waters around Australia. *J. Atmos. Ocean. Technol.* **19**, 1357–1375 (2002). doi: 10.1175/1520-0426(2002)019<1357:OIBFDW>2.0.CO;2

ACKNOWLEDGMENTS

We appreciate the commitment of the following people and sponsors: CNRS (in particular, Groupement de Recherche GDR3280); European Molecular Biology Laboratory (EMBL); Genoscope/CEA; VIB; Stazione Zoologica Anton Dohrn; UNIMIB; Fund for Scientific Research–Flanders; Rega Institute, KU Leuven; the French Ministry of Research; the French government Investissements d'Avenir programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), MEMO LIFE (ANR-10-LABX-54), PSL* Research University (ANR-11-IDEX-0001-02), and ANR (projects POSEIDON/ANR-09-BLAN- 0348, PHYTBACK/ANR-2010-1709-01, PROMETHEUS/ANR-09-PCS-GENM-217, SAMOSA/ANR-13-ADAP-0010, and TARAGIRUS/ANR-09-PCS-GENM-218); European Union FP7 (MicroB3/No.287589, IHMS/HEALTH-F4-2010-261376, and MaCuMBA/No.311975); ERC Advanced Grant Award to C.B. (Diatomite: 294823); Gordon and Betty Moore Foundation grant (no. 3790) to M.B.S.; Spanish Ministry of Science and Innovation grant CGL2011-26848/BOS MicroOcean PANGENOMICS to S.G.A.; TANIT (CONES 2010-0036) from the Agència de Gestió d´Ajusts Universitaris i Reserca to S.G.A.; JSPS KAKENHI grant no. 26430184 to H.O.; NASA Ocean Biology and Biogeochemistry program (NNX11AQ14G and NNX09AU43G) to E.B.; The Italian Research for the Sea (Flagship Project RITMARE) to D.I.; and FWO, BIO5, and Biosphere 2 to M.B.S. We also appreciate the support and commitment of Agnès b. and Etienne Bourgois, the Veolia Environment Foundation, Region Bretagne, Lorient Agglomeration, World Courier, Illumina, the EDF Foundation, FRB, the Prince Albert II de Monaco Foundation, and the *Tara* schooner and its captains and crew. We thank MERCATOR-CORIOLIS and ACRI-ST for providing daily satellite data during the expedition. We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries that graciously granted sampling permissions. *Tara* Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org). We also acknowledge excellent assistance from the European Bioinformatics Institute (EBI), in particular G. Cochrane and P. ten Hoopen, as well as the EMBL Advanced Light Microscopy Facility (ALMF), in particular R. Pepperkok. We thank Y. Timsit for stimulating scientific discussions and critical help during writing of the manuscript. The altimeter products were produced by Ssalto/Duacs and CLS, with support from CNES. The authors further declare that all data reported herein are fully and freely available from the date of publication, with no restrictions, and that all of the samples, analyses, publications, and ownership of data are free from legal entanglement or restriction of any sort by the various nations whose waters the *Tara* Oceans expedition sampled. Data described herein are available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas, at EBI under the project identifiers PRJEB402 and PRJEB7988, and at PANGAEA (*57*–*59*). The data release policy regarding future public release of *Tara* Oceans data is described in Pesant *et al*. (*50*). All authors approved the final manuscript. This article is contribution number 21 of *Tara* Oceans. The supplementary materials contain additional data. See also http://doi.pangaea.de/10.1594/PANGAEA.840721; http://doi.pangaea.de/10.1594/PANGAEA.840718; and http://doi.pangaea.de/10.1594/PANGAEA.840698

Tara Oceans Coordinators

Silvia G. Acinas,[1] Peer Bork,[2] Emmanuel Boss,[3] Chris Bowler,[4] Colomban de Vargas,[5,6] Michael Follows,[7] Gabriel Gorsky,[8,9] Nigel Grimsley,[10,11] Pascal Hingamp,[12] Daniele Iudicone,[13] Olivier Jaillon,[14,15,16] Stefanie Kandels-Lewis,[2,17] Lee Karp-Boss,[3] Eric Karsenti,[4,17] Uros Krzic,[18] Fabrice Not,[5,6] Hiroyuki Ogata,[19] Stephane Pesant,[20,21] Jeroen Raes,[22,23,24] Emmanuel G. Reynaud,[25] Christian Sardet,[26,27] Mike Sieracki,[28] Sabrina Speich,[29,30] Lars Stemmann,[8] Matthew B. Sullivan,[31] Shinichi Sunagawa,[2] Didier Velayoudon,[32] Jean Weissenbach,[14,15,16] Patrick Wincker[14,15,16]

[1]Department of Marine Biology and Oceanography, Institute of Marine Science (ICM)-CSIC, Pg. Marítim de la Barceloneta, 37-49, Barcelona E08003, Spain. [2]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [3]School of Marine Sciences, University of Maine, Orono, Maine, USA. [4]Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), and Inserm U1024, and CNRS UMR 8197, F-75005 Paris, France. [5]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [6]Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [7]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [8]CNRS, UMR 7093, LOV, Observatoire Océanologique, F-06230 Villefranche-sur-Mer, France. [9]Sorbonne Universités, UPMC Univ Paris 06, UMR 7093, LOV, Observatoire Océanologique, F-06230 Villefranche-sur-Mer, France. [10]CNRS UMR 7232, BIOM, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [11]Sorbonne Universités

Paris 06, OOB UPMC, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [12]Aix Marseille Université CNRS IGS UMR 7256, 13288 Marseille, France. [13]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [14]CEA, Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France. [15]CNRS, UMR 8030, CP5706, Evry, France. [16]Université d'Evry, UMR 8030, CP5706, Evry, France. [17]Directors' Research, European Molecular Biology Laboratory, Heidelberg, Germany. [18]Cell Biology and Biophysics, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [19]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan. [20]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany. [21]MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany. [22]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [23]Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. [24]Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [25]Earth Institute, University College Dublin, Dublin, Ireland. [26]CNRS, UMR 7009 Biodev, Observatoire Océanologique, F-06230 Villefranche-sur-Mer, France. [27]Sorbonne Universités, UPMC Univ Paris 06, UMR 7009 Biodev, F-06230 Observatoire Océanologique, Villefranche-sur-Mer, France. [28]Bigelow Laboratory for Ocean Sciences, East Boothbay, USA. [29]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond 75231 Paris Cedex 05, France. [30]Laboratoire de Physique des Océans, UBO-IUEM, Place Copernic, 29820 Plouzané, France. [31]Department of Ecology and Evolutionary Biology, University of Arizona, 1007 East Lowell Street, Tucson, AZ 85721, USA. [32]DVIP Consulting, Sèvres, France.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/1261447/suppl/DC1
Table S1

18 September 2014; accepted 23 February 2015
10.1126/science.1261447

Abstract

Diatoms (Stramenopiles, Bacillariophyceae) are an ecologically important and one of the most diverse phytoplanktonic groups, with an estimated ~1,800 marine planktonic species. Although widely studied, their diversity and biogeographic distribution patterns are not well known. The advent of high-throughput DNA sequencing has revolutionized molecular biodiversity studies facilitating the understanding of biogeography, community assembly and ecological processes. The two major goals of this thesis are (1) to investigate global biodiversity patterns and structure of marine planktonic diatom communities across the world's oceans, and (2) to understand the mechanisms and processes determining their community structure and assembly. This thesis also presents an initial attempt to discern the distribution of rare species in protist communities. The study was conducted using the metabarcoding data generated from the biological samples and associated environmental data collected during the *Tara* Oceans (2009-2013) global circumnavigation covering all major oceanic provinces. A total of ~12 million diatom V9-18S rDNA tags from 46 sampling stations, constituting 293 size fractionated samples represent the study material for the thesis. Using 63,371 unique diatom metabarcodes, this study presents an in-depth evaluation of global diatom distribution and diversity. The analyses study draw a number of revelations related to diatom biogeography, e.g. a new estimate of the total number of planktonic diatom species, a considerable unknown diversity, exceptionally high diversity in the open ocean, complex diversity patterns across oceanic provinces. The thesis then looks into the factors determining the beta-diversity patterns. The results suggest that diatoms represent biogeographically structured ecological communities regulated by both environmental heterogeneity and spatial processes. Nonetheless, the majority of the total variation in community composition remained unexplained by either the examined measured environmental factors or spatial distances, which warrants future analyses focusing on biological interactions, historical events, and other factors that are not considered. The thesis further outlines an approach to characterize significantly associated clusters of co-occurring ribotypes. Finally, a preliminary study of size-fractionated protistan communities reveals that the tail (of their rank-abundance distributions) appears to follow a power-law behavior in almost all protistan communities. This observation may indicate a potential universal mechanism which can explain the organization of marine planktonic communities. In general, this work has presented a global comprehensive perspective on diatom distribution and diversity in the world's oceans. The thesis offers an overall framework for metabarcoding-based global diversity assessments which in turn can be employed to study distribution and diversity of other taxonomic lineages. Consequently, this work provides a reference point to explore how microbial communities will respond/change in response to environmental conditions.

Résumé

Les diatomées (Stramenopiles, Bacillariophyceae) jouent un rôle important sur le plan écologique et sont l'un des groupes phytoplanctoniques les plus divers, avec environ 1800 espèces planctoniques estimées. Bien que largement étudiées, leurs modèles de diversité et de distribution biogéographique ne sont pas bien connus. L'avènement du séquençage de l'ADN à haut débit a révolutionné les études de biodiversité moléculaire facilitant la compréhension de la biogéographie, de la structure des communautés et des processus écologiques. Les deux principaux objectifs de cette thèse sont (1) d'enquêter sur les modèles de la biodiversité mondiale et la structure des communautés de diatomées planctoniques à travers les océans du monde, et (2) de comprendre les mécanismes et processus déterminants la structure de la communauté. Cette thèse présente également une première tentative de discerner la répartition des espèces rares dans les communautés de protistes. L'étude a été réalisée en utilisant les données de metabarcoding générées à partir des échantillons biologiques et des données environnementales associées recueillies au cours de la campagne *Tara* Oceans (2009-2013), une circumnavigation globale couvrant toutes les principales provinces océaniques. Le matériel d'étude pour cette thèse est constitué d'un total de 12 millions de séquences de la sous unité V9 du 18S ribosomal (barcode), récoltées à partir de 46 stations soit 293 échantillons. Basée sur 63371 metabarcodes de diatomées uniques, cette étude présente une évaluation approfondie de la distribution mondiale des diatomées et de leur diversité. Les analyses révèlent des faits marquants liées à la biogéographie des diatomées, par exemple une nouvelle estimation du nombre total d'espèces de diatomées planctoniques, une diversité considérable inconnue, une diversité exceptionnellement élevée en haute mer, et des patrons de diversité complexes entre les provinces océaniques. La thèse examine ensuite les facteurs qui déterminent les modèles de bêta-diversité. Les résultats suggèrent que les diatomées sont des communautés structurées et réglementées par l'hétérogénéité de l'environnement et des processus spatiaux. Néanmoins, la majorité de la variation totale dans la composition de la communauté ne peut être expliquée ni par les facteurs environnementaux, ni par les distances spatiales, ce qui justifie les analyses futures se concentrant sur les interactions biologiques, les événements historiques, et d'autres facteurs qui ne sont pas considérés. La thèse décrit en outre une approche pour caractériser les clusters significativement associés de ribotypes concomitants. Enfin, une étude préliminaire de communautés de protistes fractionnées par taille révèle que la queue (de leurs distributions rang abondance) semble suivre un comportement en loi de puissance dans presque toutes les communautés de protistes. Cette observation peut indiquer un mécanisme universel potentiel qui peut expliquer l'organisation de communautés planctoniques marines. De façon générale, ce travail présent une perspective globale et complète de la distribution et de la diversité des diatomées dans les océans du monde. La thèse propose un cadre global pour l'évaluation de la diversité mondiale basée sur le metabarcoding, qui pourra être utilisé pour étudier la distribution et la diversité des autres lignées taxonomiques. Par conséquent, ce travail fournit un point de référence pour explorer comment les communautés microbiennes feront face à la variation des conditions environnementales.