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

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

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

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

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

biodiversity | diatoms | metabarcoding | Tara Oceans | choke points

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

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

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

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

Significance

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

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

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

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

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

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

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

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

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

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

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

Results

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

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

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

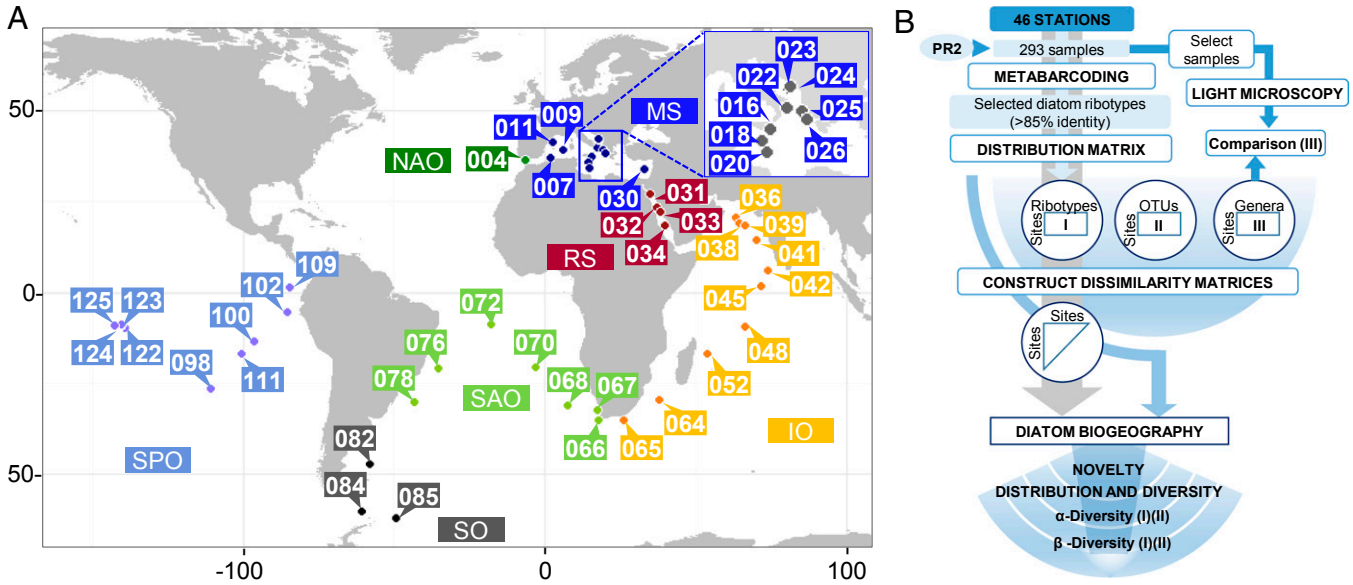

Fig. 1. Samples and methods used in the study. (*A*) Location of sampling sites (for details see ref. 37). Global diversity analysis was carried out using samples drawn from 46 global stations. At each station, the eukaryotic plankton community was sampled at two depths [subsurface (SRF) and deep chlorophyll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

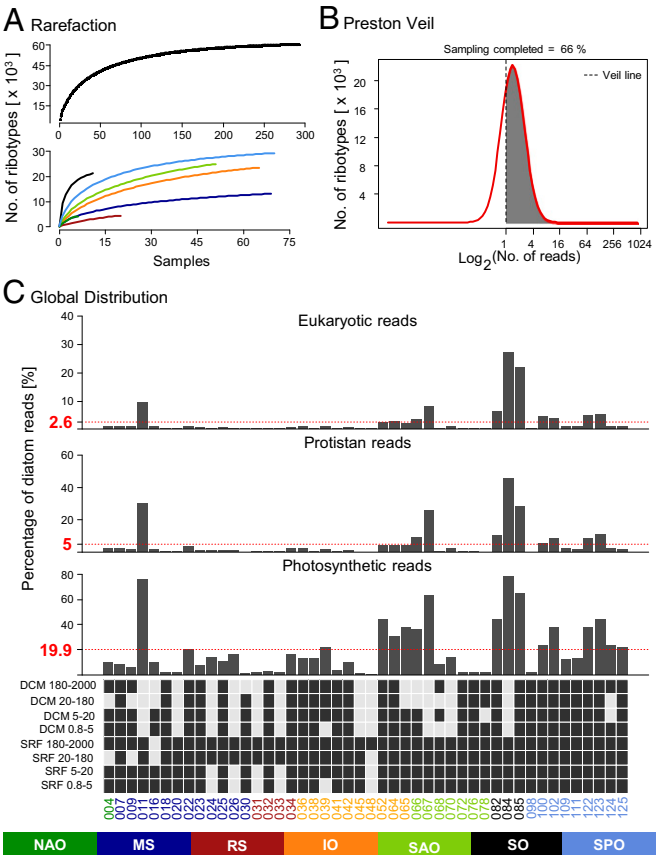

Fig. 2. Overview of the V9-rDNA diatom dataset. (*A*) V9 rDNA rarefaction curve. (*Upper*) Sample-based rarefaction curve, representing V9 rDNA richness for diatoms. (*Lower*) Each curve illustrates the estimated number of V9 rDNA sequences for each ocean province. The color code for the ocean provinces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

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

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

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

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

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

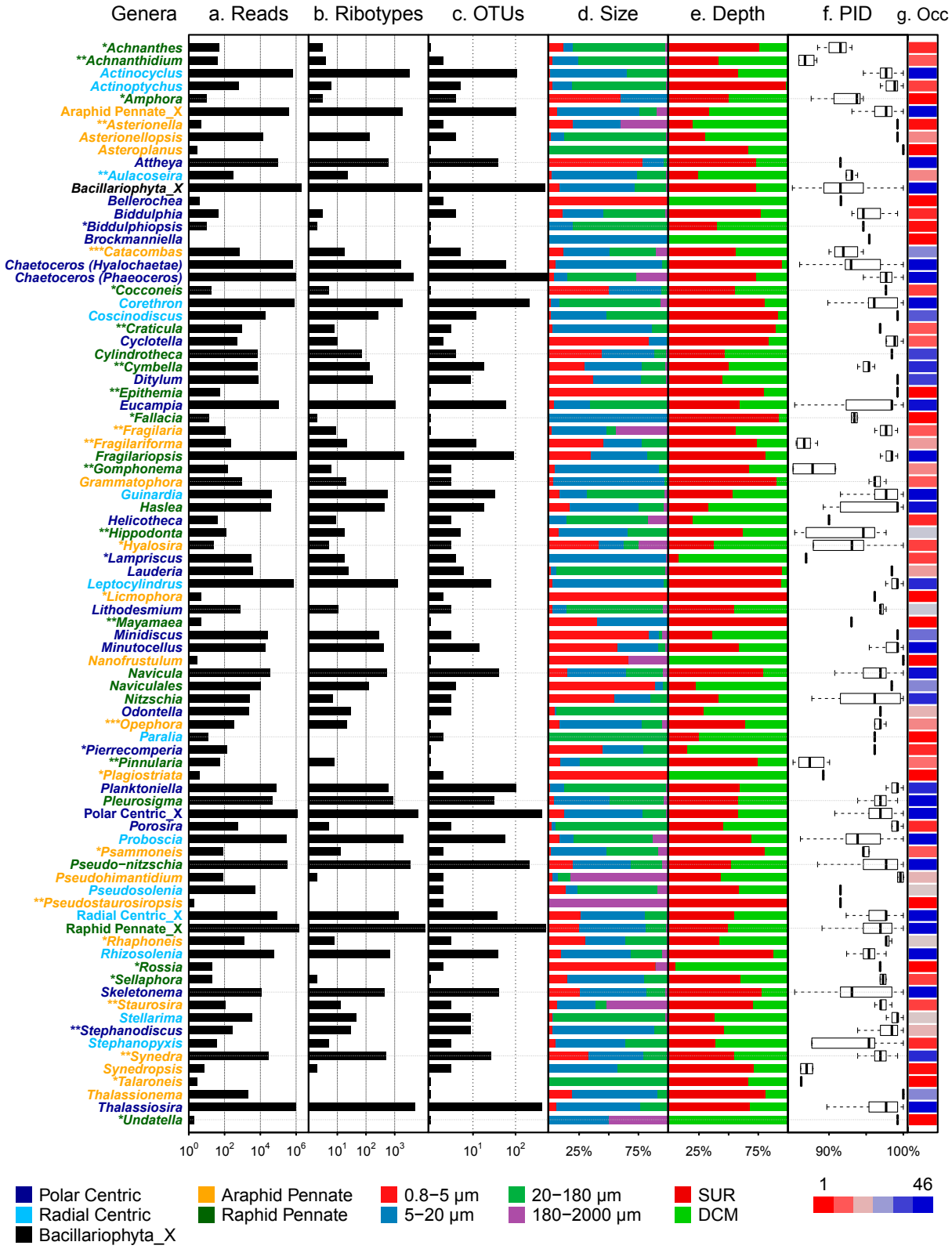

Fig. 3. Summary of diatom metabarcoding dataset. All ribotypes were clustered based on their taxonomic affiliation at the lowest taxon possible and organized under 79 genera plus five unassigned groups. The color code for a genus is as follows: dark blue, polar centric; light blue, radial centric; orange,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

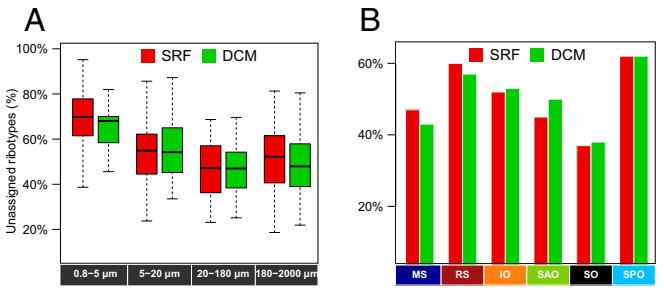

Fig. 4. Percentage of unassigned ribotypes in the *Tara* Oceans metabarcoding dataset. (*A*) Percentage of unassigned ribotypes per size class. Surface samples corresponding to 0.8–5 µm had the highest percentage of unassigned ribotypes whereas size fraction 20–180 µm had the lowest. (*B*) Percentage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

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

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

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

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

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

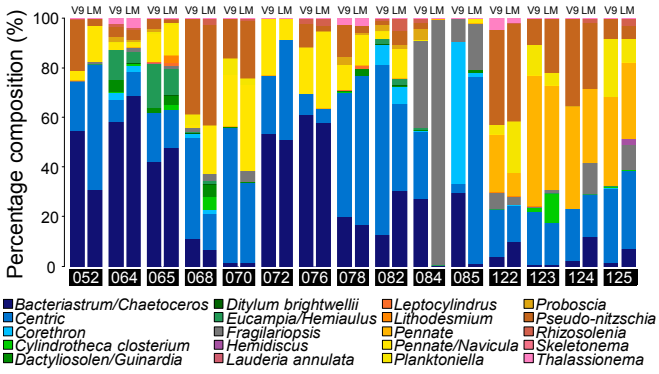

Fig. 5. Comparisons of diatom community compositions estimated from V9 rDNA counts and by light microscopy (LM). Shown are community composition profiles obtained from light microscopy and ribotype relative abundance inferred from taxonomy-based clustering of assigned ribotypes from 15 selected stations.

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

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

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

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

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

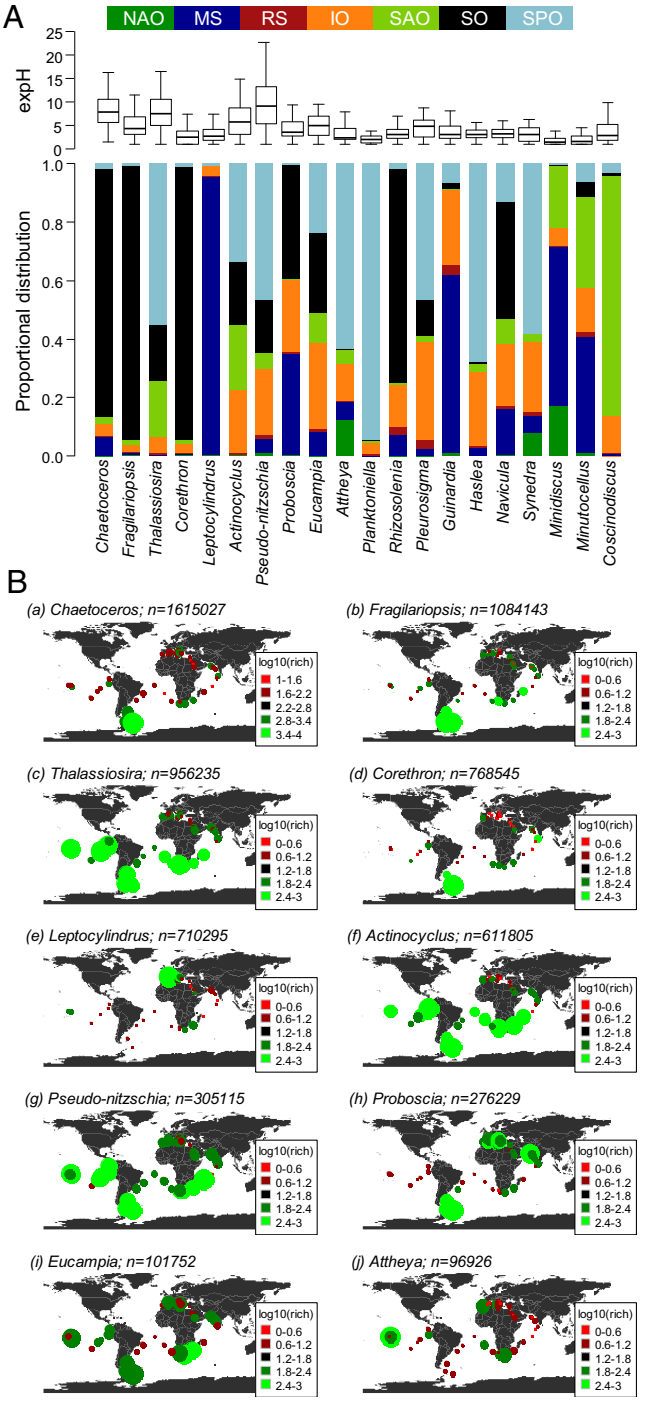

Fig. 6. Local and regional genus distribution and diversity inferred from *Tara* Oceans dataset. (*A*) Distribution of top 20 diatom genera in seven oceanic provinces. These genera accounted for 98.84% of the assigned reads in the entire dataset. (*Upper*) The variation in diversity for each indicated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

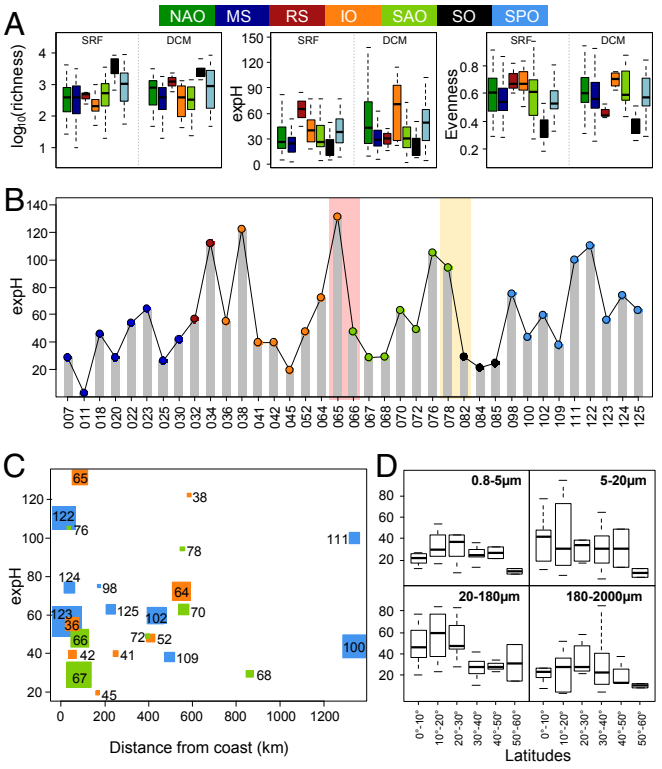

Fig. 7. Variation in diatom diversity across oceanic basins. (*A*) Variation in richness (expressed as number of unique ribotypes), diversity [expressed as exponentiated Shannon diversity index (expH)], and evenness across provinces. (*B*) Variation in diatom diversity across 37 stations (expH) for which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

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

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

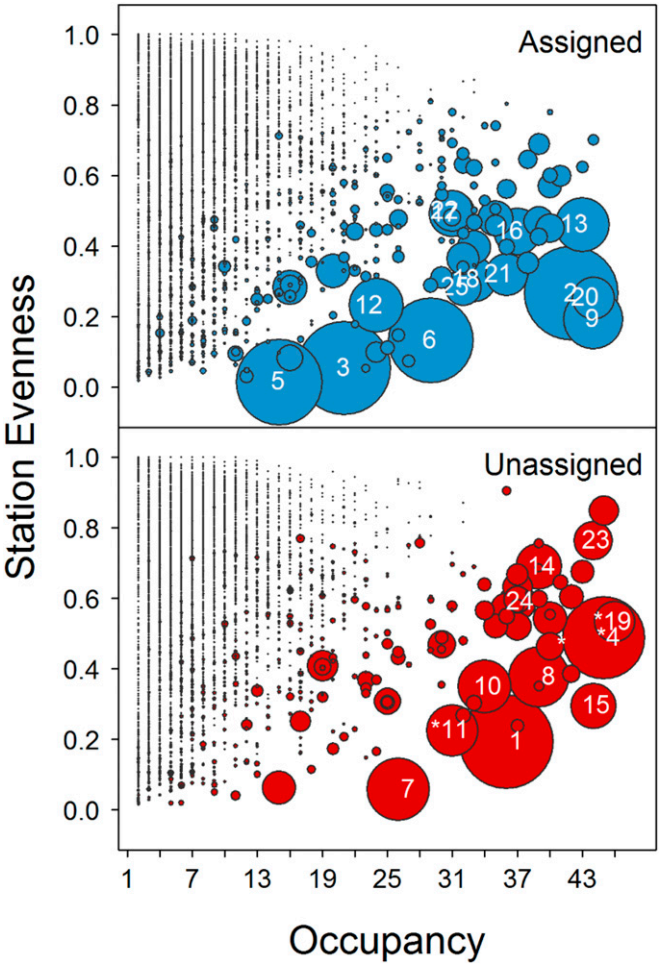

Fig. 8. Cosmopolitanism, total abundance, and station evenness of each diatom ribotype. (*Upper*) Ribotypes that could be assigned to a genus/species. (*Lower*) Ribotypes that could not be assigned to any genus. Each circle represents a ribotype (V9 rDNA), the radius being scaled to the number of reads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

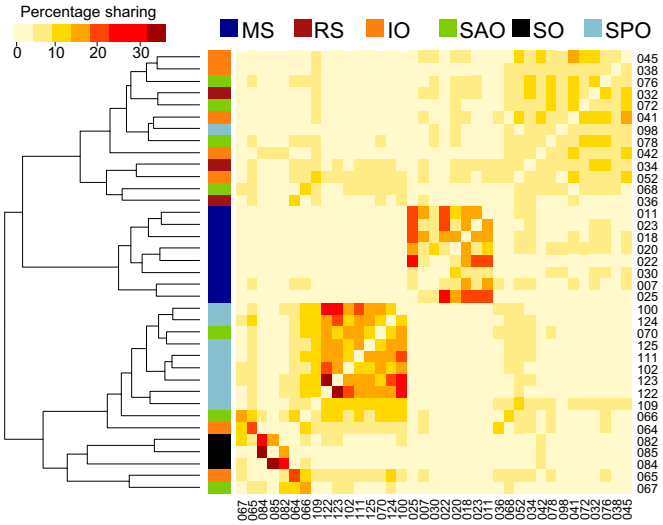

Fig. 9. Biogeographic patterns. Percentage of ribotypes shared between stations. Only those stations for which surface samples for all size classes were available are shown (37 stations). For each station, a pooled community derived from all size classes was obtained. A dendrogram of complete linkage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

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

Discussion

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

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

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

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

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

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

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

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

Materials and Methods

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

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

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

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

1. Smetacek V (1998) Diatoms and the silicate factor. *Nature* 391:224–225.
2. Falkowski PG (2002) The ocean's invisible forest. *Sci Am* 287(2):54–61.
3. Armbrust EV (2009) The life of diatoms in the world's oceans. *Nature* 459(7244): 185–192.
4. Nelson DM, Treguer P, Brzezinski MA, Leynaert A, Queguiner B (1995) Production and dissolution of biogenic silica in the ocean: Revised global estimates, comparison with regional data and relationship to biogenic sedimentation. *Global Biogeochem Cycles* 9:359–372.

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