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

Shruti Malviya,a1, Eleonora Scalo,b Stéphane Audic,c Flora Vincent,a Alagaruj Veluchamy,a,2 Julie Poulain,d Patrick Wincker,d,e,f Daniele Iudicone,a Colombian de Vargas,g Lucie Bittner,a,3 Adriana Zingone,b and Chris Bowlera,4


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 Fragilaria, 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.

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 the 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.

Significance

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” (9) 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, 11). 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, and metabarcoding have explored the dataset generated by the 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.

Supporting Information

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.

1Present address: Biological Oceanography Division, National Institute of Oceanography, Dona Paula, Goa 403 004, India.
2Present address: Biological and Environmental Sciences and Engineering Division, Center for Desert Agriculture, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia.
3Present address: Sorbonne Universités, Université Pierre et Marie Curie (UPMC), CNRS, Institut de Biologie Paris-Seine (IBPS), Evolution Paris Seine, F-75005 Paris, France.
4To 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.
Identification methods using light microscopy (LM) were done on a number of selected samples to validate the molecular data. 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 protist 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. 2A) 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. 2B), suggesting that our survey retrieved ~66% of diatom ribosomal diversity in the photic zone of the global ocean (shaded region in Fig. 2B). 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.](https://www.pnas.org/cgi/doi/10.1073/pnas.1509523113) 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.
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. 2C 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 *Thalassiothrix* (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 diatom groups (the global oceans’ main phytoplankton) at time intervals of several decades (31). 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*, *Urospora*) or benthic and marine littoral species (e.g., *Amphiprora*, *Caloneis*, *Ardisoneis*, *Hyalodiscus*, *Pseudoisthia*, *Entomoneis*, *Phaeodactylum*), except for only a few pelagic marine genera (e.g., *Bacillaria*) (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. L4). 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. S4D).

Intragenus diversity was found to vary from as low as one ribotype per genus (e.g., *Nanofrustulum*, *Asteroiplanusa*, *Belleroca*) 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 *Hylochoa* (*Dataset S1*). *Chaetoceros* (both subgenera), *Thalassiothrix*, *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., *Aithya*, *Ditylum*, and *Belleroca*) (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., *Hyalosina*, *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*, *Aithya*, 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.
Asterotheca, Bellerochea, Helicotheca, Nanoconfusum, and Lithodendron were seen mostly in deep chlorophyll maximum (DCM) samples (Fig. 3, column c). The level of percentage identity to the reference sequence also varied across genera (Fig. 3, column f). Pseudo-nitzschia, Actinocyclus, Attheya, Chaetoceros, Eucampia, Fragilariopsis, Minamitoriella, 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. 4A). 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. 4B).

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 Hyalochaete) and Fragilariposis 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 Asterolaem, Asteromphalus, Climacodium, Dactyliosolen, Hemiaulis, Hemiaulis, and Ladera.

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 Planctoniella were among the least diverse and that this observation also reflected the known differences in species richness for these genera (Fig. 6A). Most diatom genera were seen in all oceanic provinces although their abundance patterns were highly variable: for instance, Chaetoceros (both subgenera), Corethron, and Fragilariposis were highly abundant in the SO, in accordance with previous data (e.g., ref. 43); Attheya, Planctoniella, 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. 6B and SI Appendix, Fig. S6). Three main patterns were found, with some genera having a lower diversity in the tropics (e.g., Fragilariposis, Probosica, and Eucampia), others showing lower diversity at high latitudes (e.g., Attheya, Guinardiella), and others with a more uniform diversity (e.g., Thalassiosira, 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 environments. It is noteworthy that GenBank annotations were available for only a small subset of the genera having a high-
areas (SI Appendix, Fig. S6D). 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. S6C). 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. 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 also observed (SI Appendix, Fig. S7A). 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. S7A). 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. 7B and SI Appendix, Fig. S7B). Diversity was significantly lower in the DCM samples from the Maldives (station 45, North IO), but increased toward the north and the south (Fig. 7B and SI Appendix, Fig. S7B). Station 11 in the MS displayed the lowest diversity of all, the result of a diatom bloom that was dominated by *Leptocylindrus* (Figs. 1C, 6B, and 7B). 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. 2C), we found no significant difference in the diversity at open ocean stations versus coastal stations (Fig. 7C). Indeed, diversity showed no correlation with diatom *W* diversity 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. 7D), 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 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 biculatus* ("**4**"), *Asteromphalus spp. ("**11**"), *Pseudo-nitzschia delicatissima* ("**19**") and *Thalassithrix longissima* ("**4**") (SI Appendix, SI Materials and Methods). Several 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. *Pit for* 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).
The actual SPO subset is made up of tropical stations in an HNLC, coherent with their relative homogeneity of conditions (for instance, and MS showed the highest degree of internal similarity (Fig. 9), size fractions were available (37 stations). Stations in the SPO, SO, considering the similarity among surface stations for which all four provinces (Fig. S7), ribotypes (out of 63,371; 0.9%) were present in all oceanic basins, whereas only 576 mostly because of the coastal SAO stations; the SAO/IO (8,569 is downstream of the former; the SAO and SPO (9,501 ribotypes), where the latter particular, the SPO and IO (12,176 ribotypes), where the latter.

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 SPO and SO (9,501 ribotypes), mostly because of the coastal SAO stations; the SAO/IO (8,569 ribotypes); and the SO/IOO (7,330 ribotypes)) whereas only 576 ribotypes (out of 63,371; 0.9%) were present in all oceanic provinces (SI Appendix, Fig. S7D). 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, Fragilariapopsis; 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, 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”, “**11”- Asteromphalus spp., “**19”- Pseudo-nitzschia delicatissima, and “**2”- Thalassiothrix longissima (SI Appendix, SI Materials and Methods).
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 fully resolve these temporal differences.

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 Actinoptychus (∼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 Actinoptychus 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, Actinoptychus, Pseudo-nitzschia, Proboscia, and Eucampia showed almost equal worldwide distributions across all provinces (in agreement with ref. 62). In general we found consistency in our 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 Koçolek (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 assignment 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 lower (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 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 retroflexion 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 (DNA metabarcoded ribotypes) were generated by the CMA method from the 293 communities. Please see de Vargas et al. (36) for details on the sequencing and taxonomic assignment 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 araphid 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. Piélon’s evenness (75) and the exponentiated Shannon–Weiner 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’Énergie 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 GENOMIQUE (ANR-10-INBS-09-08), POSEIDON (ANR-09-BLAN-0348), PROMETHEUS (ANR-09-GENM-031) and RITMARE project. By a grant from the Ministero dell'Istruzione dell'Avenur—programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (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 (Micro3No.287358), European Research Council Advanced Grant Award (to C.B.) (Diatomite: 294283), Agnès b., the Veolia Environment Foundation, Region Bretagne, World Courier, Illumina, Cap L’Orient, the Électricité de France (EDF) Foundation EDF Diversité, Fondation pour la Recherche sur la Biodiversité, the Prince Albert II de Monaco Foundation, Etienne Bourgois, and the Eléctricité de France (EDF) Foundation EDF Diversité, 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.


Supporting Information

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

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

This file includes:

Materials and Methods
Figures S1 to S8
Dataset 1 and 2

S1 Materials and Methods
Distance based Analysis

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

Extracting diatom V9 metabarcodes from global eukaryotic protistan metabarcoding data set

A total of ~580 million quality-checked reads, representing ~2.3 million unique rDNA ribotypes (V9 region of 18S rDNA), were generated from 334 photic-zone plankton communities sampled during the Tara Oceans expedition (3). The Tara Oceans expedition (4-5) covered seven oceanographic provinces, i.e., North Atlantic Ocean (NAO), Mediterranean Sea (MS), Red Sea (RS), Indian Ocean (IO), South Atlantic Ocean (SAO), Southern Ocean (SO), and South Pacific Ocean (SPO). At each station, plankton communities were obtained for four size fractions from two water-column depths (SRF and DCM). Total nucleic acids (DNA + RNA) were extracted from all samples, and the hyper-variable V9 region of the nuclear 18S rDNA was PCR-amplified (3). The V9 reads were quality checked and to reduce the influence of PCR and sequencing errors, only sequences seen in at least two different samples with at least 3 copies were retained. The sequences have been deposited in GenBank (see [4], Accession: PRJEB4352; ID: 213098). Taxonomy assignments for all ribotypes were obtained through annotation against an expert-curated V9 reference database (for details, see [3]) using the global alignment search strategy implemented in the ggsearch36 program (Fasta package). This reference database contains sequences from both cultured strains and the environment, and contained 1,232 unique diatom V9 reference sequences corresponding to 159 genera, with most genera being represented by more than one sequence (Fig. S2). Of the 159 genera in the reference data set, we retrieved 87 genera in our data set. However, only 79 out of 87 were assigned at an identity greater than 85% and were selected for further analysis.
These unique barcodes were taxonomically assigned to known eukaryotic entities based on the PR2 database (1). From this, metabarcodes assigned to diatoms, at a percentage identity of ≥ 85% to the reference sequence, were selected. When BLAST results gave rise to more than one unique best hit, a last common taxonomy of the best BLAST hits was created [3]. Moreover, in order to improve the assignment of the barcodes that couldn’t be assigned to the genus level, as the PR2 database version we used was based on a former release of Genbank [3], PR2 assignments were also compared to Genbank assignments (release 210 from October 2015). We manually checked each assignment, and kept the best of two (PR2 or Genbank assignment) based on the best percentage identity value and BLAST scores. We could thus improve our conclusions and assignments, in particular when working with sequences that could not be assigned to the genus level (SI Dataset 2).

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

**Morphological analyses**

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

**Reassignment of unknown diatom ribotypes**

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

For each target diatom ribotype, the equivalent 18s rDNA preamplified sample in which its relative abundance was the highest was chosen for PCR (Polymerase Chain Reaction), in order to maximise chances of amplifying the ribotype with highly specific reverse primers. The forward primer chosen was the D512F (D512F: 5’-ccgcttaacctgataagcttggcgcgt3’) universal
diatom primer from Zimmerman et al. (8). The reverse primer was designed in order to find the
3’ end consecutive eight base pairs 100% specific to the target sequence that matched the lowest
number of non-specific sequences in the sample. Four ribotypes and their respective reverse
primer sequences are listed below:

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Ribotype md5sum ID</th>
<th>Reverse primer sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV9_172</td>
<td>01eb4d181204cc0e142f5f5f1632b0b8c</td>
<td>172_rev: 5’-aggtcggacaagtctcgcggtcag-3’</td>
</tr>
<tr>
<td>TV9_225</td>
<td>4d2d2df1f3cd2b038ace0b23c17928be</td>
<td>225_rev: 5’-ttctctactaaatgataaggttagacagt -3’</td>
</tr>
<tr>
<td>TV9_361</td>
<td>ba6e7a54f4f24e0888797d4e062cda61</td>
<td>361_rev: 5’-ggggacaagttctcgegctaacat-3’</td>
</tr>
<tr>
<td>TV9_339</td>
<td>8e6521a0e823436605e0d302c33da9</td>
<td>339_rev: 5’-gcggagacaagttctcgcagagat-3’</td>
</tr>
</tbody>
</table>

TV9_179: St82[0.8-inf]; TV9_225: St85[20-180]; TV9_361: St123[5-20]; TV9_339: St122[5-20]
dem

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

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

References
cellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids
Res 41(D1):D597-D604
348(6237):1261605.
data. Sci Data 2:150023.


Assessing V9 hypervariable sub-sequence (130 bp) of small-subunit (SSU) ribosomal RNA (rRNA) genes as diversity markers. (A) 2,947 full-length 18S rDNA sequences were obtained from the PR2 reference database corresponding to 718 diatom species (1). 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. 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. Length variation and pairwise genetic distances calculated using the Kimura-2-parameter model for all nine hypervariable regions are shown. Regression of V1-V9 p-distance by NJ on to that of the 18S
sequence shows that V5 could best explain the phylogeny, followed by V4. Although the mean genetic distances were better in V4 and V9, they may not explain the phylogeny well. The performance of V9 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. (D) Phylogenetic inference on Bacillariophyta from full-length 18S rDNA sequence phylogeny, V4 rDNA phylogeny, and V9 rDNA based phylogeny rooted at Stramenopiles. 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 MEGA6. The tree was statistically tested using 1,000 bootstraps. 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).
Fig. S2. Novelty in *Tara* Oceans diatom ribotype data set. Barplot showing the number of reference sequences present for each genus and the total number of unique V9 tags from *Tara* Oceans data set 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, 30,057 ribotypes were unassigned due to the lack of reference sequence.
Fig. S3. Completeness of the diatom ribotype data set based on OTUs. (A) OTU rarefaction curve. A sample-based rarefaction curve, representing OTU richness for diatoms. (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 4,748, indicating 873 OTUs undetected. OTU calling was based on Mahé et al. (6).
Fig. S4. Comparing diatom distributions and abundance obtained from our study to the reports from MAREDAT and OBIS. (A) Venn diagram showing the overlap between Tara Oceans data set, the V9 PR2 reference database, and MAREDAT. The green circle represents the subset of reference genera identified in the Tara Oceans data set. (B) Distribution of Bacillariophyta obtained from MAREDAT database (9). The colour represents log-transformed cell counts per litre. (C) Distribution of Bacillariophyta obtained from Tara Oceans data set. The colour represents log-transformed total V9 ribotype abundance. (D) Global abundance of Bacillariophyta species obtained from OBIS datasets (10) (each square is coloured according to the abundance of diatoms species observed in the area).
Fig. S5. Percentage of unassigned ribotypes in each station. Within each station, 31-81% of the ribotypes could not be assigned to any known diatom genera. The highest proportion of unassigned ribotypes was seen in Station 45 (~79%) followed by stations in the Pacific Ocean (Stations 109, 111, 122, 123, 124) (~58-68%). Stations with the highest abundance (Stations 67 and 85) contained ~34% of unassigned ribotypes.
A

(a) Planktoniella; n=81044

(b) Rhizosolenia; n=54766

(c) Pleurosiga; n=48954

(d) Guinardia; n=41831

(e) Haslea; n=36856

(f) Navicula; n=33561

(g) Synedra; n=28700

(h) Minidiscus; n=25960

(i) Minutocellus; n=18283

(j) Coscinodiscus; n=18038

B

(a) Chaetoceros (Phaeoceros); n=962087

(b) Chaetoceros (Hyalochaetae); n=652940
Fig. S6. Global distribution and diversity of abundant genera. (A) Genera ranked 11 to 20 based on their ribotype abundance. (B) Chaetoceros (Phaeoceros) and Chaetoceros (Hyalochaetae). (A & B). The area of the circle is scaled to the total number of reads for each genus at each location. For each panel, the color key is shown in the legend. Red - low richness; Green - high richness. (C) Abundance of the two Chaetoceros subgenera in different stations. (D) Abundance in different stations of each Chaetoceros subgenus compared to distance from the coast. Each square corresponds to a station. Left panel was built using reads and right panel was built using ribotypes. Red dots/squares correspond to Hyalochaetae and blue dots/squares correspond to Phaeoceros data.
Fig. S7. Variations in diatom diversity. (A) Ribotype richness, effective number of species (expressed as expH) and evenness are shown per depth and size-class. The results indicate that the 20-180 µm size fraction was the most diverse and showed highest diversity at DCM. The smaller 0.8-5 µm size fraction also showed similar trends. The largest size fraction exhibited the lowest abundance and richness but has the highest evenness among all size fractions. In general, SRF samples were found to be less diverse and even than DCM samples. (B) Variation in diatom diversity across stations. Spatial variation of diatom diversity across 37 stations inferred from Shannon Diversity Index (SDI) and richness. (C) Pairwise community dissimilarity (Bray-Curtis) across provinces, signifying higher dissimilarities for higher values. (D) Shared number of ribotypes among oceans. Bar graph showing the number of ribotypes shared between oceanic provinces; 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.
Fig. S8. Diatom community composition. (A) Incidence measure. Pairwise Jaccard dissimilarity was used to cluster stations hierarchically (group-average linkage). A two-dimensional NMDS ordination indicated that communities grouped according to oceanic provinces, albeit with significant overlapping (stress=0.19). Each of these oceanic clusters were significantly different (ANOSIM; R = 0.58; W = 0.001). (B) Abundance-based measure. 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 oceanic province.