

Supplementary Figures and Tables

Patterns and drivers of diatom diversity and abundance in the global ocean

Pierella Karlusich, et al. 2025 *Nat Comm*

Table of contents

Table S1: Estimated ASV richness for V4 and V9 datasets using multiple biodiversity estimators.

Table S2: Partial least square analyses to explore the correlations between diatom relative abundance and Shannon index with the physicochemical context.

Table S3: Partial least square analyses to explore the correlations between diatom relative abundance and biotic factors.

Figure S1: Biogeographical coverage of the *Tara* Oceans datasets used in the current work.

Figure S2: Samples and methods used in this study.

Figure S3: Contribution of diatoms to eukaryotic phytoplankton and silicifiers in surface waters of the global ocean using V4 metabarcoding data obtained from different size-fractions.

Figure S4: Accumulation curves for V4 and V9 18S rRNA gene metabarcoding assigned to diatoms.

Figure S5: Distribution of diatom relative abundances in epipelagic seawater samples from the V9 and V4 metabarcoding data across size fractions, water layers, ocean basins, and biomes.

Figure S6: Comparison between V4 and V9 metabarcoding for the global macroecological patterns of diatom relative abundance and diversity in surface waters using data obtained from different size-fractions.

Figure S7: Abundance patterns of diatom classes based on the V4 marker.

Figure S8: Relative contribution of the different diatom classes.

Figure S9: Biogeography of diatom classes in surface waters using V4 and V9 metabarcoding data obtained from different size-fractionated samples.

Figure S10: Environmental distribution of the *Attheya* genus.

Figure S11: Global distribution of the 20 most abundant genera in the V9 and V4 datasets.

Figure S12: Taxonomic composition of the diatom ASV modules obtained by WGCNA.

Figure S13: Unassigned ASVs in the *Tara* Oceans V4 and V9 metabarcoding datasets from diatoms.

Figure S14: Phylogenetic distribution of DUF285 (PF03382) in reference proteomes from UniprotKB database.

Figure S15: Latitudinal abundance gradient for genes and transcripts coding for Domain of Unknown Function 285 (DUF285) and Heat Shock Protein 90 (HSP90) in diatoms and other eukaryotic phytoplankton.

Table S1: Estimated ASV richness for V4 and V9 datasets using multiple biodiversity estimators.

Marker	Observed ASV count	Chao1 estimator	Chao1 standard error	First-order Jackknife estimator	First-order Jackknife standard error	Second-order Jackknife estimator	Bootstrap estimator	Bootstrap standard error	Total number of observations
V9	34243424	3424.899	1.04073 9	3441.981	4.238084	3280.522	3472.61	25.21701	931
V4	3429	3429.878	1.04383 4	3443.985	4.791743	3331.35	3473.272	16.97236	969

Table S2: Partial least square analyses to explore the correlations between diatom relative abundance and Shannon index with the physicochemical context

	<i>Diatom relative abundance</i>		<i>Diatom Shannon</i>	
	<i>VIP score</i>	<i>Standard coefficient</i>	<i>VIP score</i>	<i>Standard coefficient</i>
<i>Temperature</i>	1.6266765279	-0.18715260	1.5764379	0.14211408
<i>NH₄⁺/DIN</i>	0.7449139294	-0.06770999	0.7251692	0.04639623
<i>NH₄⁺</i>	0.0004108426	-0.04073561	0.3672286	0.04230855
<i>NO₂⁻</i>	0.0186188635	-0.05158923	0.4473117	0.05296705
<i>NO₃⁻</i>	1.1938164317	0.10643676	1.1647857	-0.07219932
<i>Fe</i>	0.0761344935	0.04829419	0.3701818	-0.04770001
<i>Si</i>	1.0040927560	0.09568034	0.9730387	-0.06711979
<i>PO₄³⁻</i>	0.8829044972	0.05801240	0.9078752	-0.03189882
<i>Chlorophyll a</i>	1.1668409466	0.12469494	1.1261470	-0.09202232
<i>Absolute latitude</i>	1.4894446409	0.18268462	1.4563720	-0.14187917

Table S3: Partial least square analyses to explore the correlations between diatom relative abundance and biotic factors

	<i>Diatom relative abundance</i>	
	<i>VIP score</i>	<i>Standard coefficient</i>
<i>Prochlorococcus</i>	2.0589894	-0.236453116
<i>Synechococcus</i>	0.7559680	-0.087359284
<i>Copepoda</i>	1.4889173	0.151032888
<i>Centroheliiozoa</i>	1.5517588	-0.059139034
<i>Choanoflagellida</i>	1.7055238	0.195750218
<i>Chrysophyceae</i>	1.4418286	-0.007121123
<i>Dictyochophyceae</i>	2.3466947	-0.082131732
<i>Nassellaria</i>	1.4064072	-0.162014670
<i>Phaeodarea</i>	0.4485323	-0.051816101
<i>Spumellaria</i>	0.9581452	-0.094281817
<i>Rhizaria</i>	0.5690121	-0.055823984

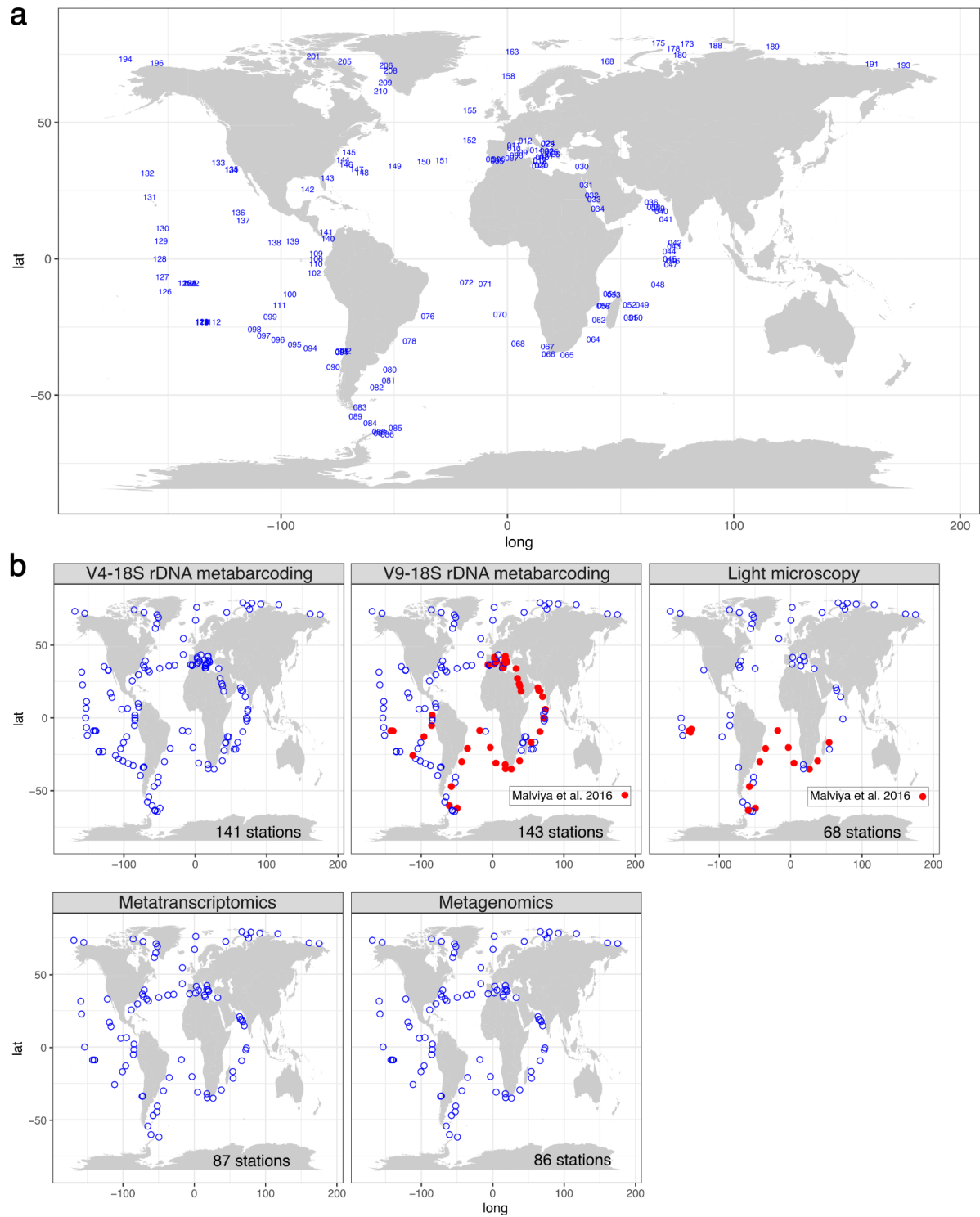


Figure S1: Biogeographical coverage of the *Tara Oceans* datasets used in the current work. a) Sampling station labels. b) Biogeography of the datasets. The total number of stations are indicated. The sampling stations reported by Malviya et al. 2016 *PNAS* 113 (11):E1516-E1525 are marked in red: 43 sampling stations for the V9-based survey and 15 stations for the microscopy survey. Maps were generated with the *borders()* function in *ggplot2*⁸⁴.

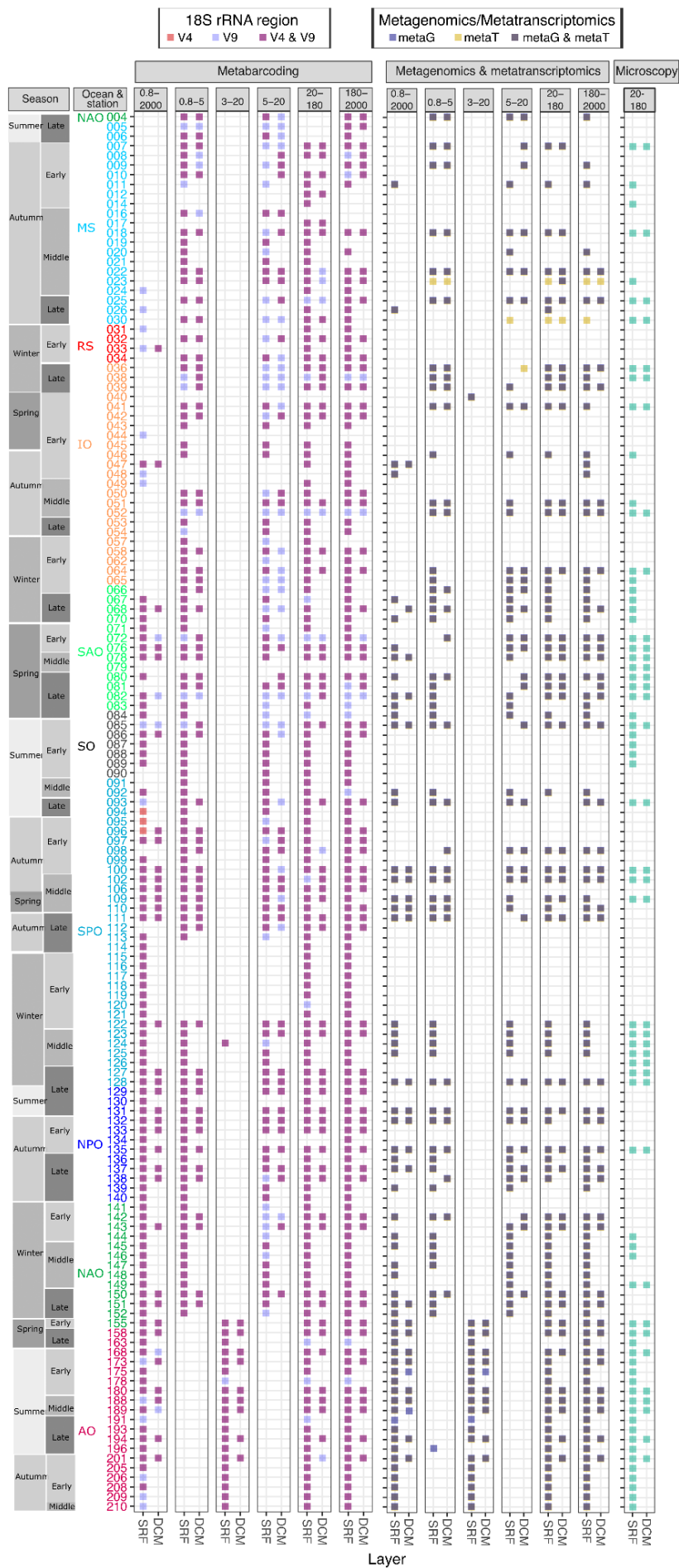


Figure S2: Samples and methods used in this study. The current analysis of global diversity and abundance of diatoms was carried out across 145 *Tara* Oceans stations where samples were taken for metabarcoding and/or metagenomics and/or metatranscriptomics and/or optical microscopy. The analyzed samples are indicated as filled boxes, with color according to the method. Two distinct depth layers were sampled: surface (SUR; 5 m) and deep chlorophyll maximum (DCM; 17-180 m). The data from the bottom of the mixed layer was collected when no DCM was observed (stations TARA_123, TARA_124, TARA_125, TARA_152 and TARA_153). Plankton communities were fractionated into four size classes: piconanoplankton (0.8 to 5 μm or 0.8 to 2000 μm), nanoplankton (5 to 20 μm or 3 to 20 μm), microplankton (20 to 180 μm), and mesoplankton (180 to 2000 μm). Season and moment of the season (early, middle, late) are displayed to the left of the panel. Station labels are coloured according to the ocean region: IO, Indian Ocean; MS, Mediterranean Sea; NAO, North Atlantic Ocean; RS, Red Sea; SAO, South Atlantic Ocean; SO, Southern Ocean; SPO, South Pacific Ocean.

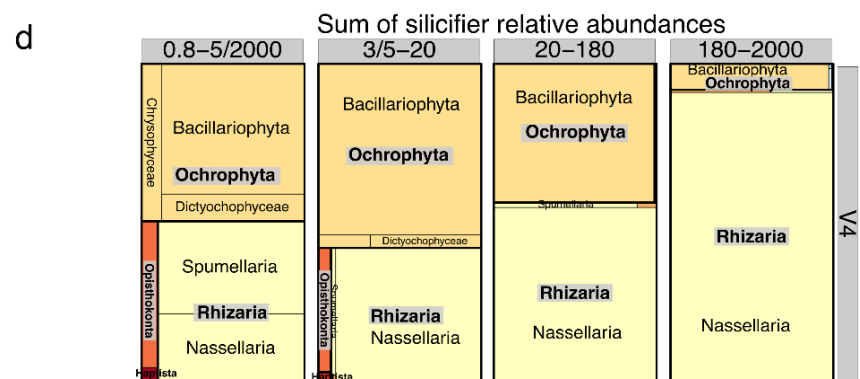
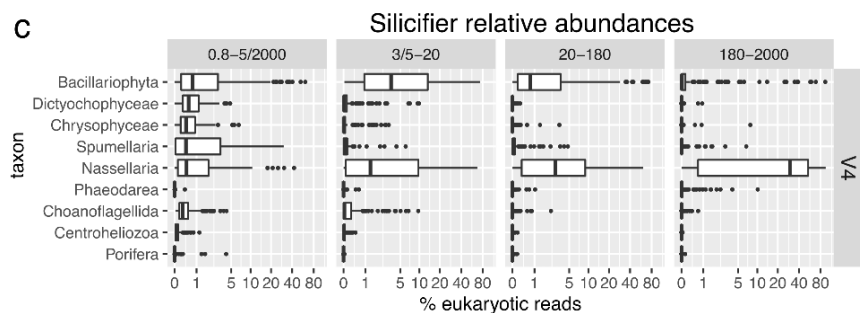
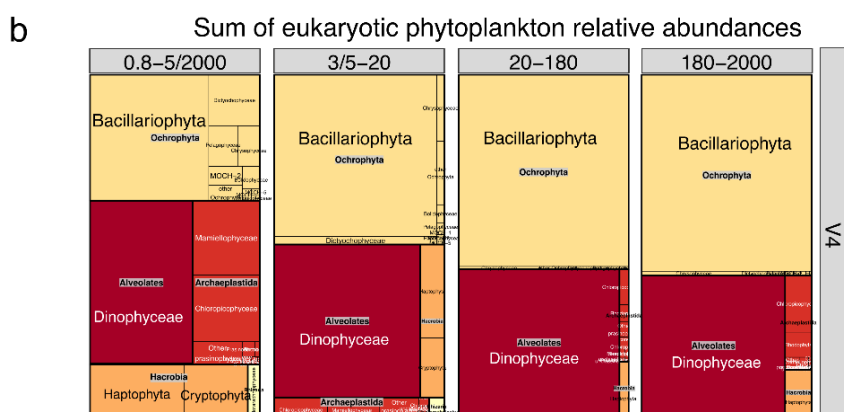
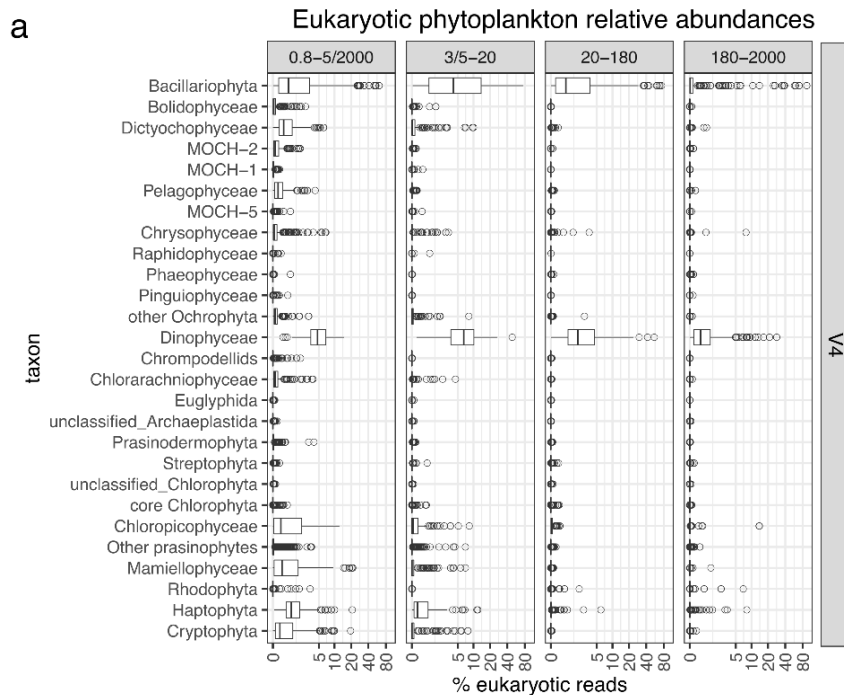


Figure S3: Contribution of diatoms to eukaryotic phytoplankton (a-b) and silicifiers (c-d) in surface waters of the global ocean using V4 metabarcoding data obtained from different size-fractions. a-b) Phytoplankton. We focused exclusively on the phytoplankton signal of these data sets, including dinoflagellates and chrysophytes though we acknowledge there are uncertainties in assigning the capacity for photosynthesis in these groups. a) Relative abundances (log scale). Each point is a size-fractionated sample. b) Sum of normalized reads in the overall dataset. c-d) Silicifiers. The equivalent plots for the V9 marker are displayed in Fig 1. Boxplots illustrate the distribution of the dataset, with the box representing the 25–75% interquartile range and the central line indicating the median (50% quantile). Whiskers extend to data points within 1.5 times the interquartile range. The V4 dataset comprises 184 samples for the 0.8–5 μm or 0.8–2000 μm size fractions, 127 for the 3–20 μm or 5–20 μm fractions, 175 for the 20–180 μm fraction, and 185 for the 180–2000 μm fraction.

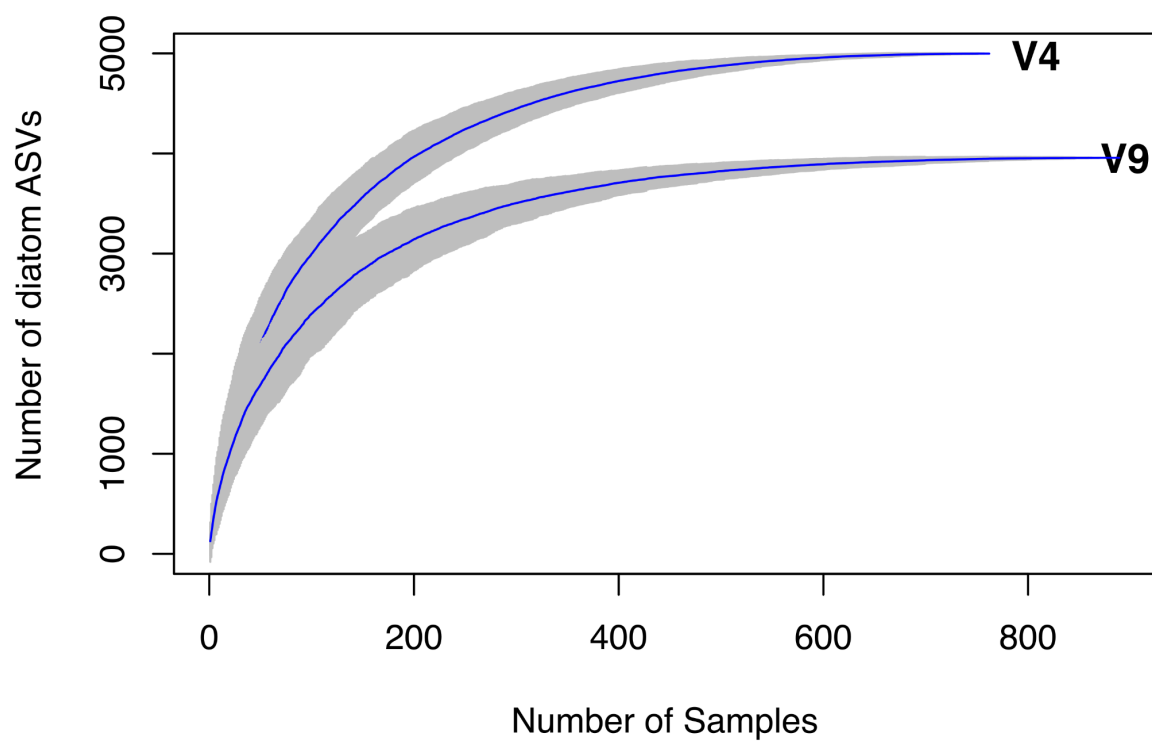


Figure S4: Accumulation curves for V4 and V9 18S rRNA gene metabarcoding assigned to diatoms.

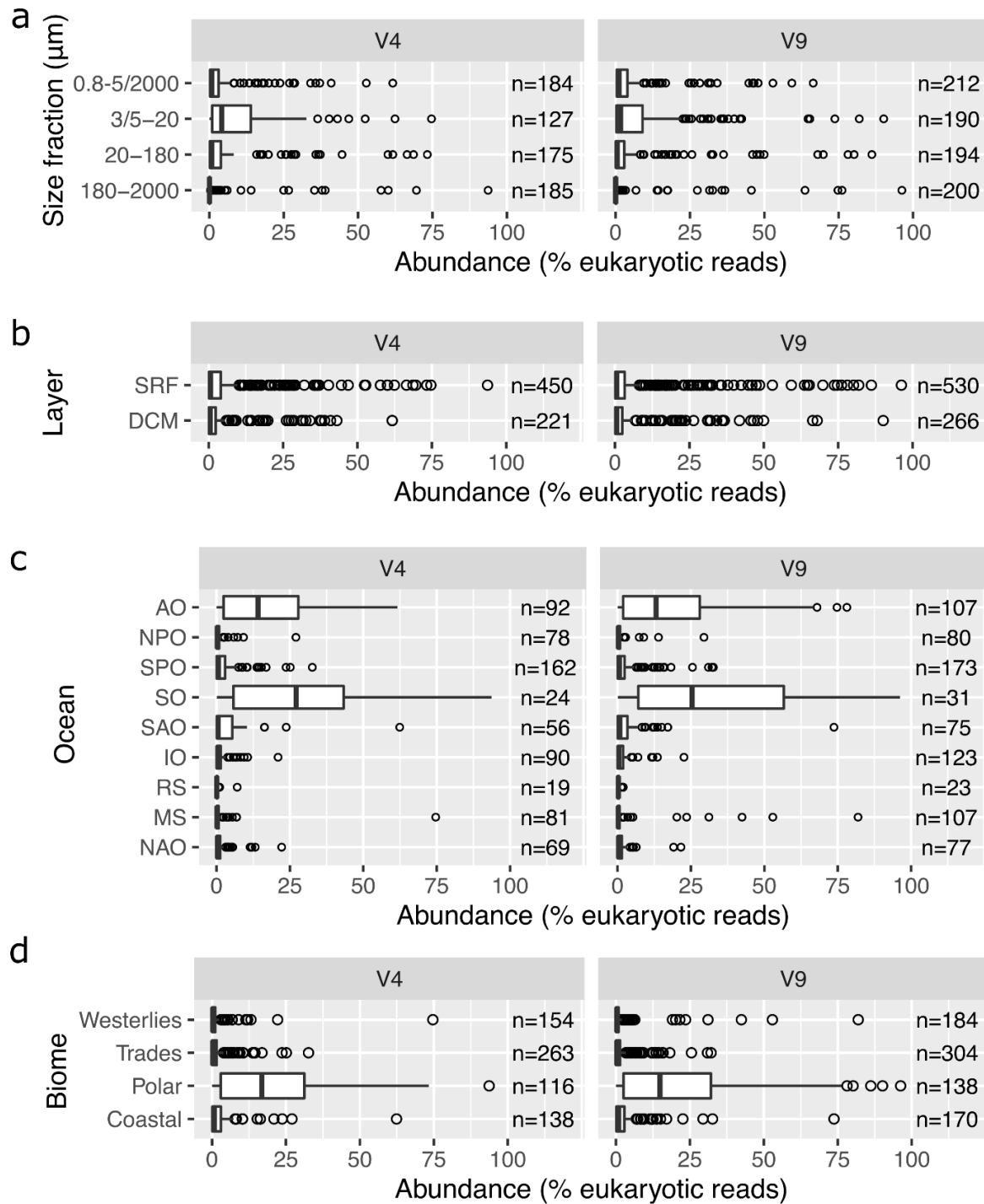


Figure S5: Distribution of diatom relative abundances in epipelagic seawater samples from the V9 and V4 metabarcoding data across size fractions, water layers, and ocean basins. a) Size fractions. b) Water layers (SRF, surface; DCM, deep chlorophyll maximum). c) Ocean basins (MS, Mediterranean Sea, IO, Indian Ocean, SAO, South Atlantic Ocean, SO, Southern Ocean, SPO, South Pacific Ocean, NPO, North Pacific Ocean, NAO, North Atlantic Ocean, AO, Arctic Ocean). d) Biomes.

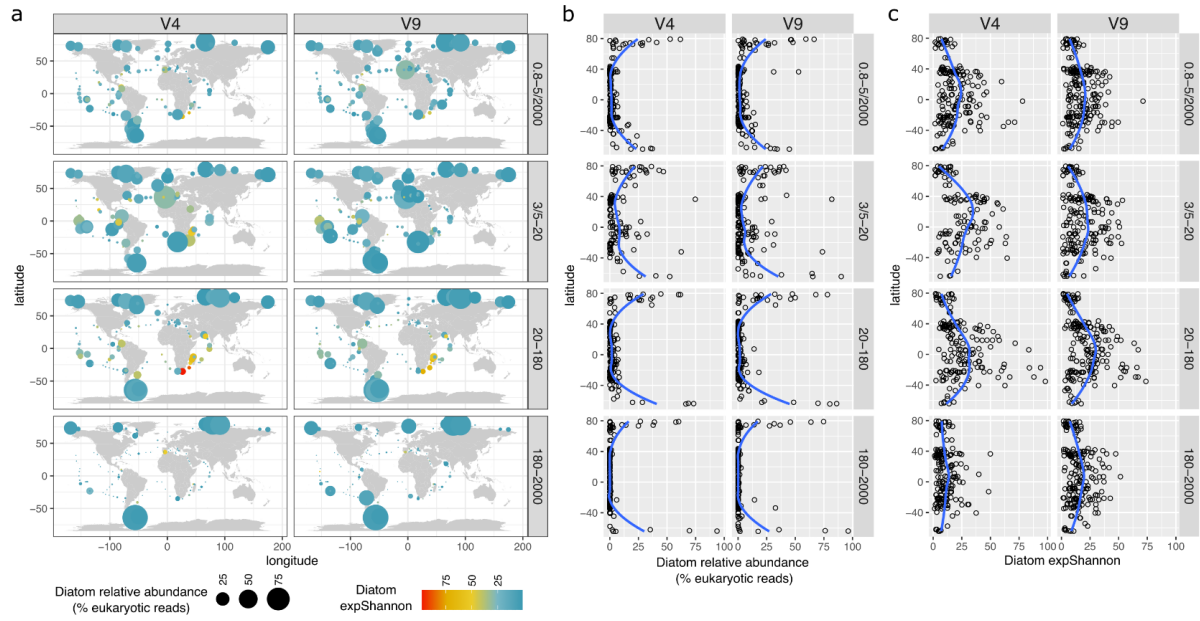


Figure S6: Comparison between V4 and V9 metabarcoding for the global macroecological patterns of diatom relative abundance and diversity in surface waters using data obtained from different size-fractions. a) Distribution maps. b) Latitudinal gradient for relative abundance. c) Latitudinal gradient for the exponentiated Shannon Diversity Index. The blue lines correspond to Loess smoothings. Maps were generated with the *borders()* function in *ggplot2*⁸⁴.

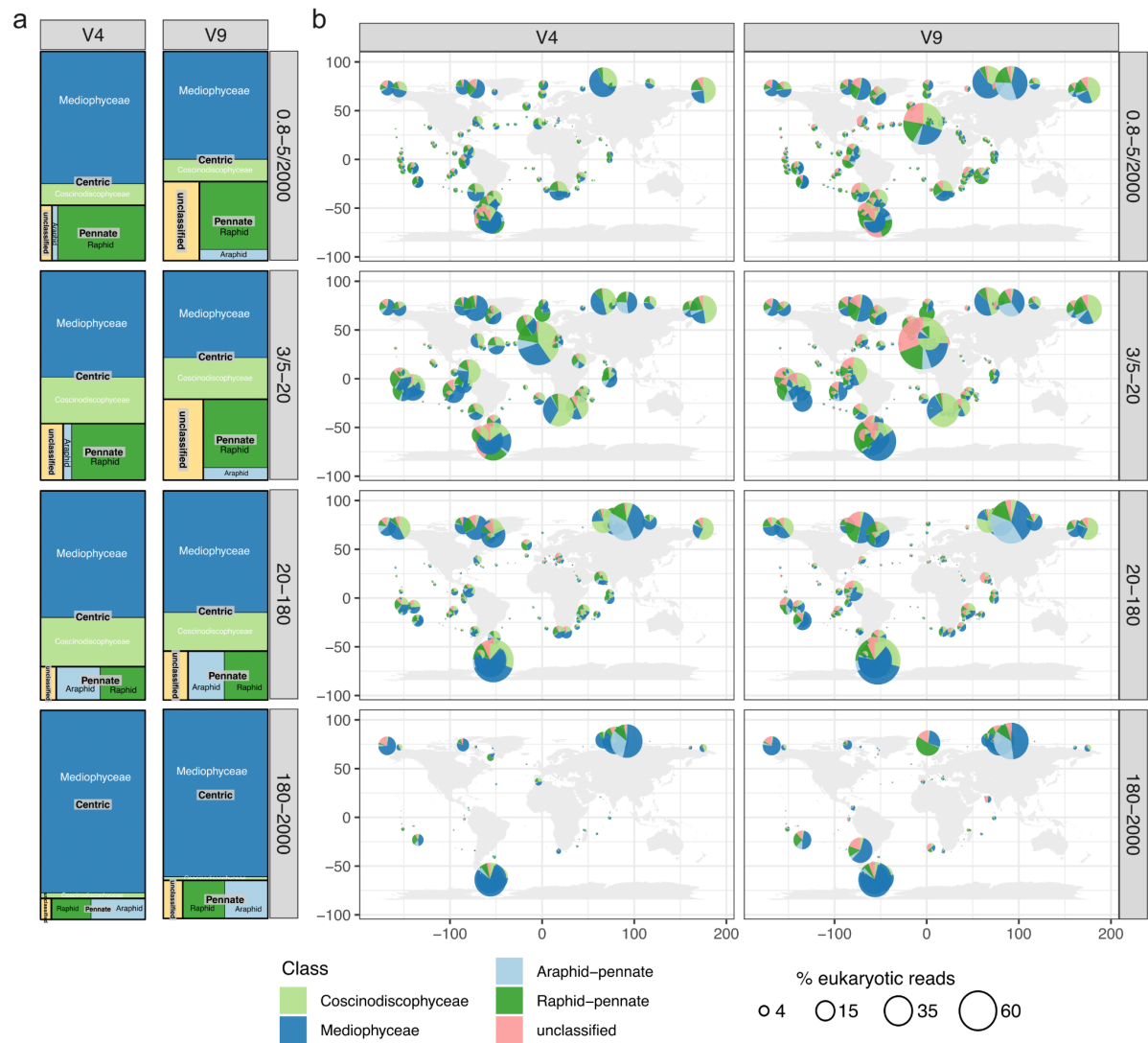


Figure S8: Relative contribution of the different diatom classes. a) Sum of normalized reads for V4 and V9 markers in the overall dataset. b) Biogeography. The maps separated by diatom classes are shown in Fig S9. Maps were generated with the *borders()* function in *ggplot2*⁸⁴.

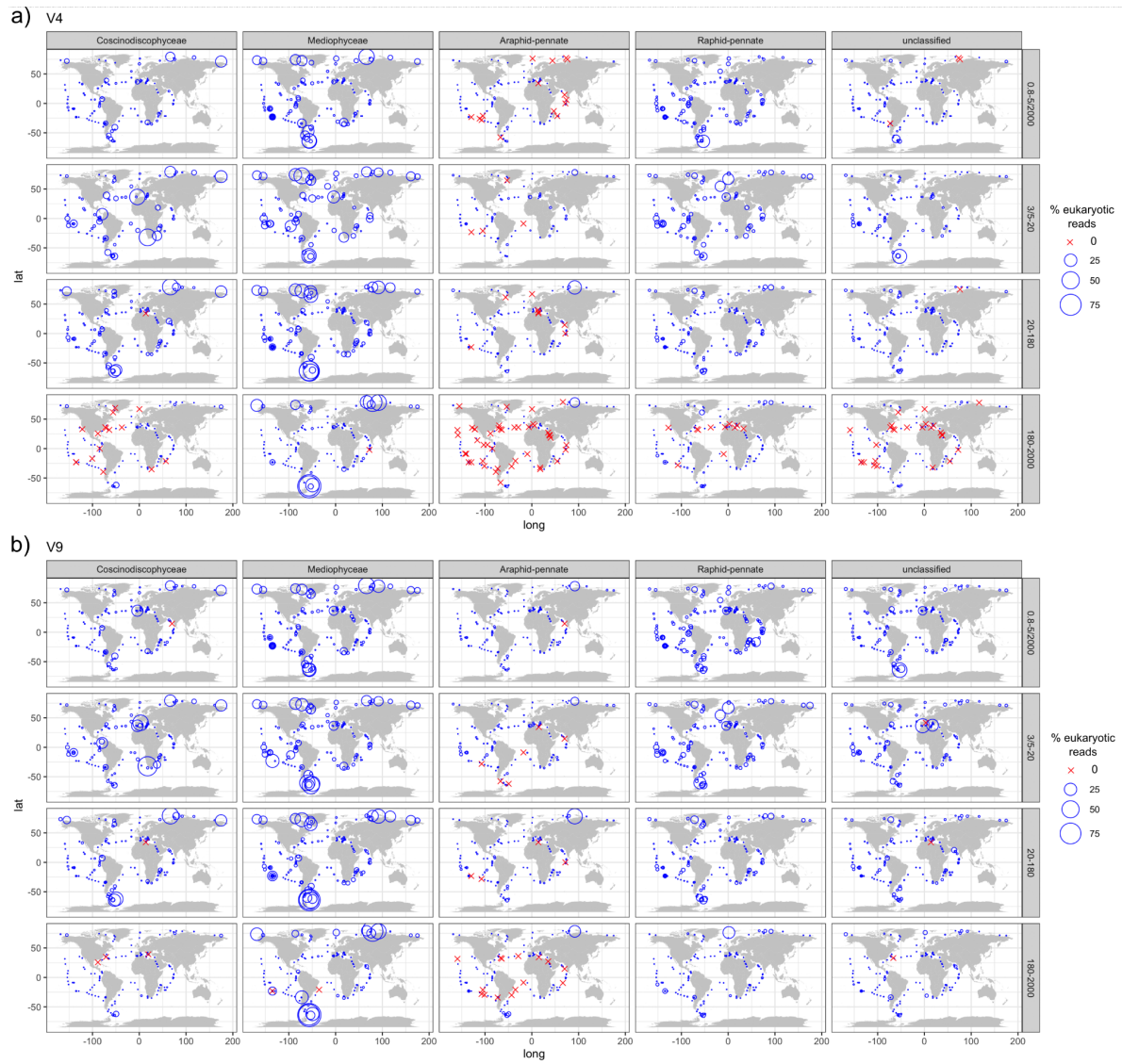


Figure S9: Biogeography of diatom classes in surface waters using V4 (a) and V9 (b) metabarcoding data obtained from different size-fractionated samples. The bubble sizes vary according to the percentage of diatoms among eukaryotes. Maps were generated with the *borders()* function in *ggplot2*⁸⁴.

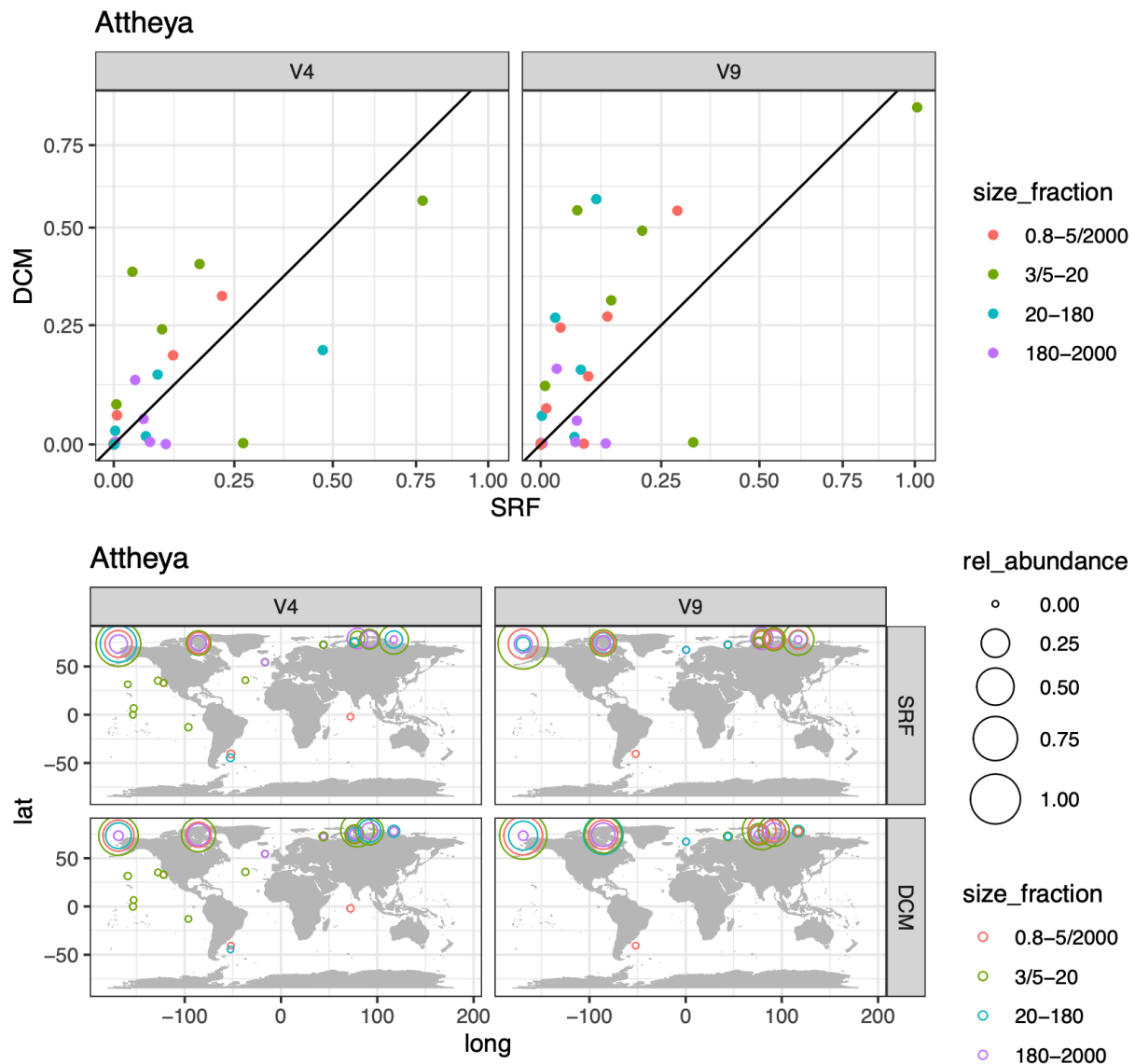


Figure S10: Environmental distribution of the *Attheya* genus. a) Depth distribution. The scatter plots compare the relative abundances between surface (5 m) and deep chlorophyll maximum (DCM; 17–180 m). Axes are in the same scale and the diagonal line corresponds to a 1:1 slope. b) Biogeography. The bubble sizes vary according to the percentage of diatoms among eukaryotes. The data corresponds to V4 (left panel) and V9 (right panel) metabarcoding data obtained from different size-fractionated samples (indicated in colors). Maps were generated with the *borders()* function in *ggplot2*⁸⁴.

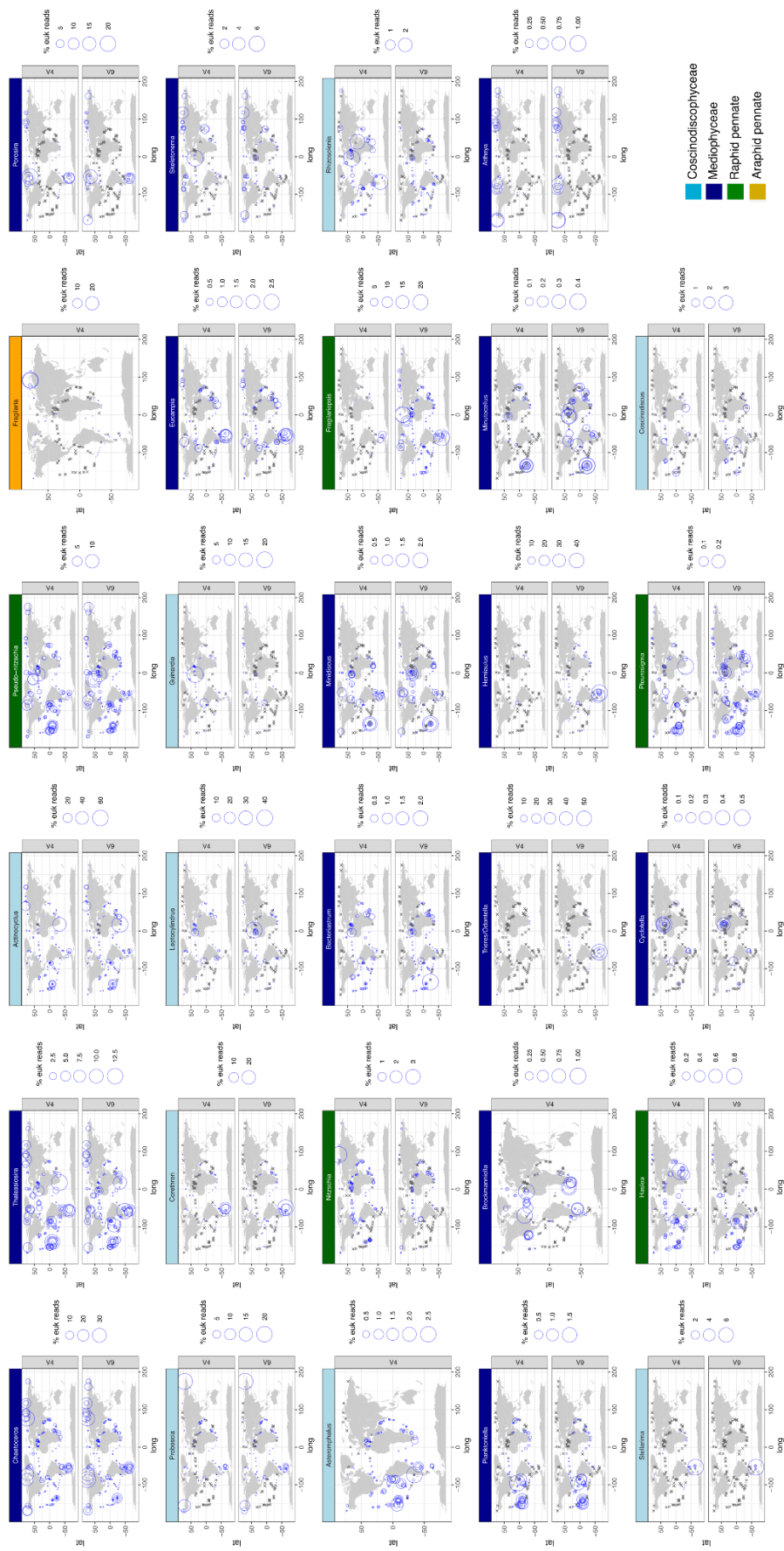


Figure S11: Global distribution of the 20 most abundant genera in the V9 and V4 datasets. All genera in the list of Fig 6a are displayed. This figure is analogous to Figure 7, but differentiates by marker region. Bubble areas represent the percentage of reads for each genus among eukaryotic reads at each station location, with crosses indicating absence of detection. The maps depict surface samples from the size fraction where the genus was most prevalent: 20-180 μm for *Chaetoceros*, *Fragilaria*, *Porosira*, *Proboscia*, *Corethron*, *Eucampia*, *Asteromphalus*, *Planktoniella* and *Trieres/Odontella*, and *Stellarima*; 3/5-20 μm for *Thalassiosira*, *Actinocyclus*, *Pseudo-nitzschia*, *Guinardia*, *Leptocylindrus*, *Fragilariopsis*, *Skeletonema*, *Bacteriastrum*, *Rhizosolenia*, *Attheya*, *Haslea*, *Pleurosigma*, *Coscinodiscus*, and *Hemiaulus*; 0.8-5/2000 μm for *Nitzschia*, *Minidiscus*, *Brockmanniella*, *Minutocellus*, and *Cyclotella*. The maps including all size fractions and covering the top 50 most abundant genera are available in Supplementary File 1. Maps were generated with the *borders()* function in *ggplot2*⁸⁴.

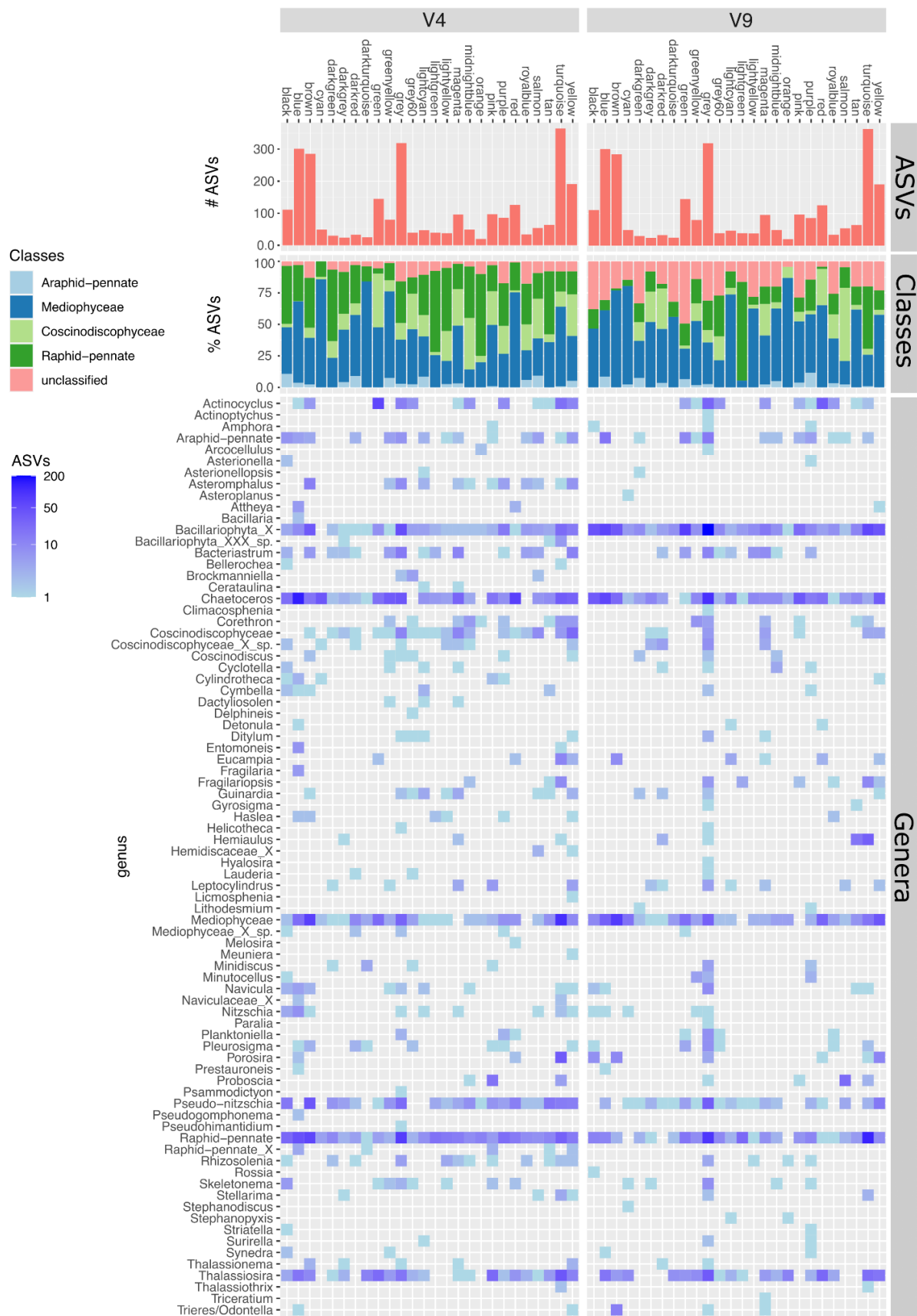


Figure S12: Taxonomic composition of the diatom ASV modules obtained by WGCNA.

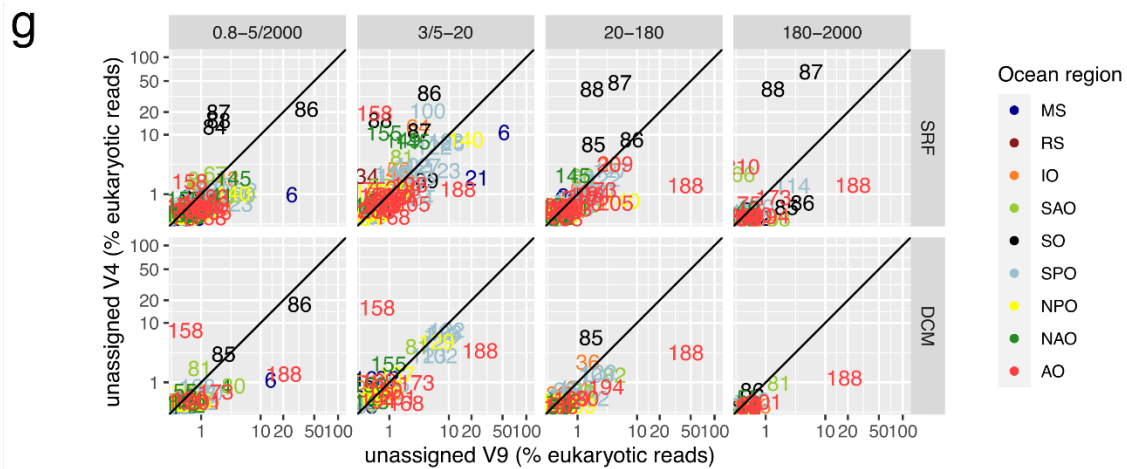
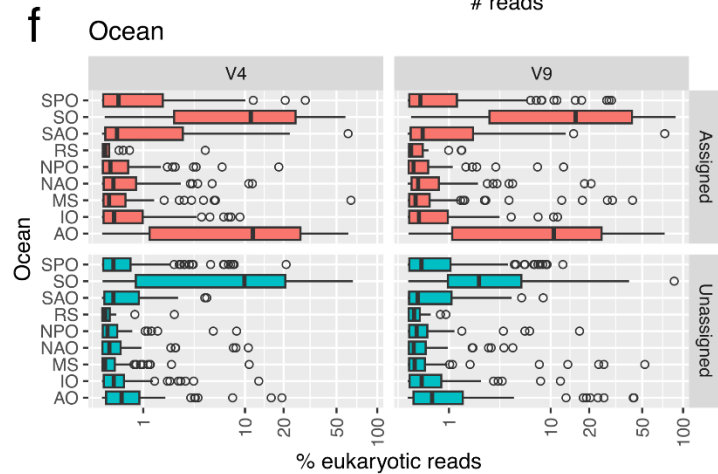
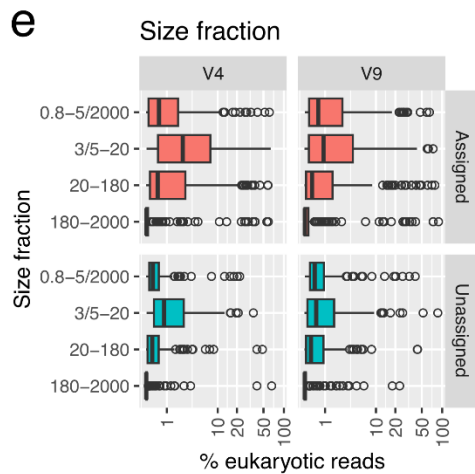
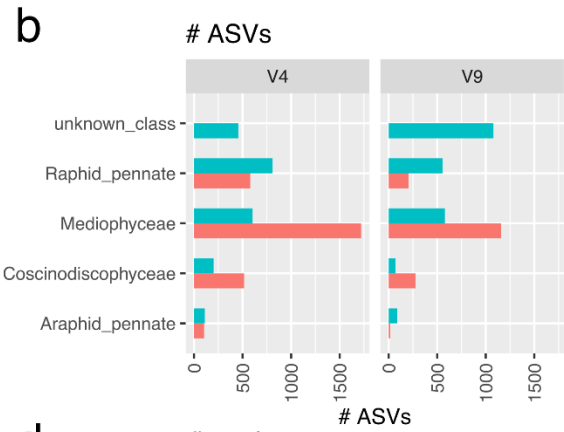
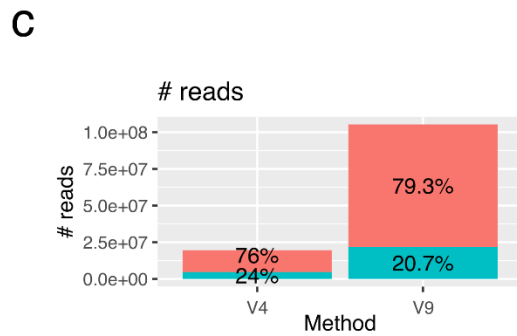
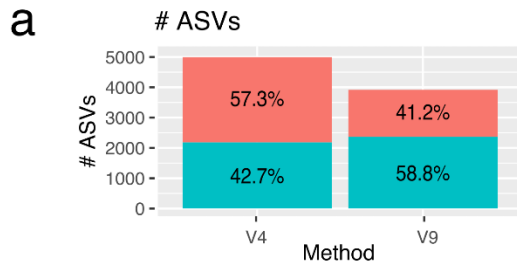
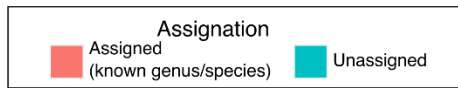


Figure S13: Unassigned ASVs in the *Tara* Oceans V4 and V9 metabarcoding datasets from diatoms. The ASVs considered to be unassigned were those that could not be unambiguously assigned to any diatom genus/species but could be classified only as araphid or raphid pennate, Mediophyceae, Coscinodiscophyceae, or unassigned diatom. a-b) Total number of assigned (red) and unassigned (blue) ASVs, and their class-level distribution. c-d) Total number of assigned (red) and unassigned (blue) reads, and their class-level distribution. e) Read abundances of assigned and unassigned ASVs per size class. f) Read abundances of assigned and unassigned ASVs by ocean region. g) Comparison of the read abundance of unassigned ASVs in those samples that are represented by both V4 and V9 metabarcoding. Axes are in the same logarithmic scale and the diagonal line corresponds to a 1:1 slope. Station labels are indicated, with the color according to the ocean region: MS Mediterranean Sea, IO Indian Ocean, SAO South Atlantic Ocean, SO Southern Ocean, SPO South Pacific Ocean, NPO North Pacific Ocean, NAO North Atlantic Ocean, AO Arctic Ocean. Each point in e-g panels corresponds to a size-fractionated sample.

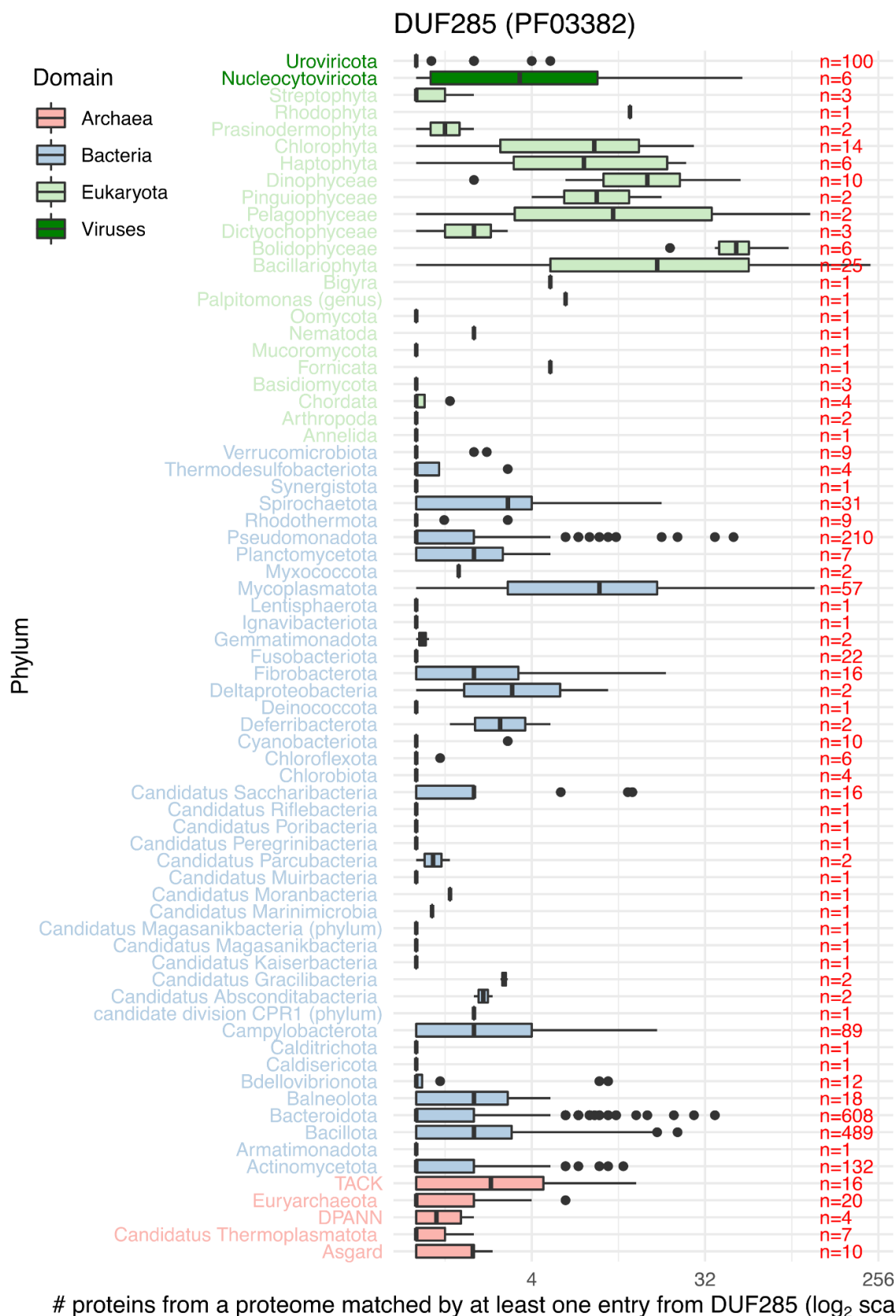


Figure S14: Phylogenetic distribution of DUF285 (PF03382) in reference proteomes from UniprotKB database (<https://www.uniprot.org/uniprotkb>). The number of analysed species is displayed in red. The x axis corresponds to the number of unique DUF285 sequences per proteome (\log_2 scale)

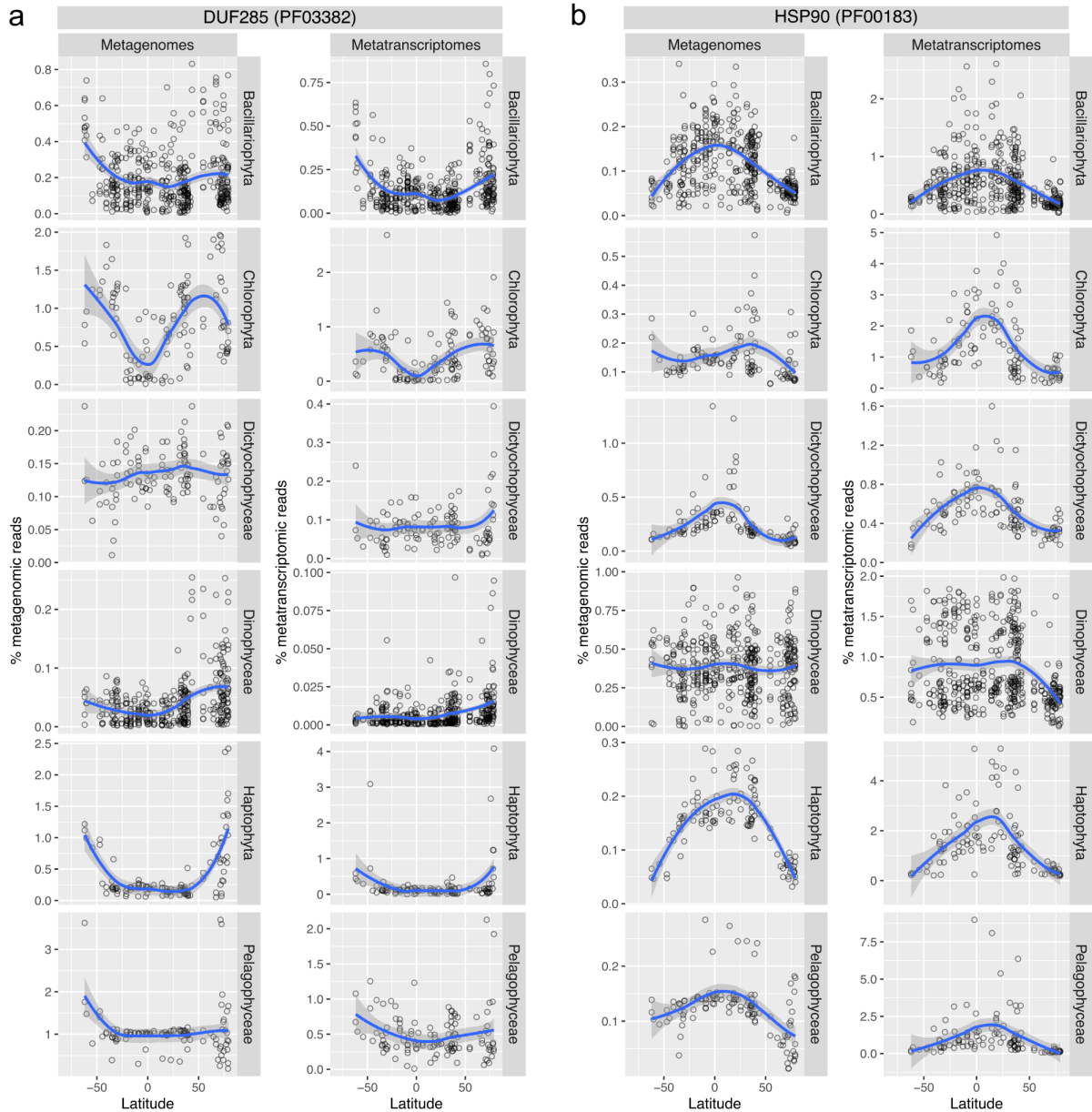


Figure S15: Latitudinal abundance gradient for genes and transcripts coding for Domain of Unknown Function 285 (DUF285) and Heat Shock Protein 90 (HSP90) in diatoms and other eukaryotic phytoplankton. a) DUF285 (PF03382). b) HSP90 (PF00183). The scatter plots correspond to samples from the size fraction where the taxon was most prevalent: 0.8-5/2000, 3/5-20, 20-180, and 180-2000 μm for diatoms and dinoflagellates, and 0.8-5/2000 μm for the rest. Y axis corresponds to the proportion of metagenomic or metatranscriptomic reads for the given function among the total metagenomic or metatranscriptomic reads from the corresponding taxon. Note that we were not able to discard heterotrophic species from dinoflagellates due to the small number of reference gene catalogs for this group.

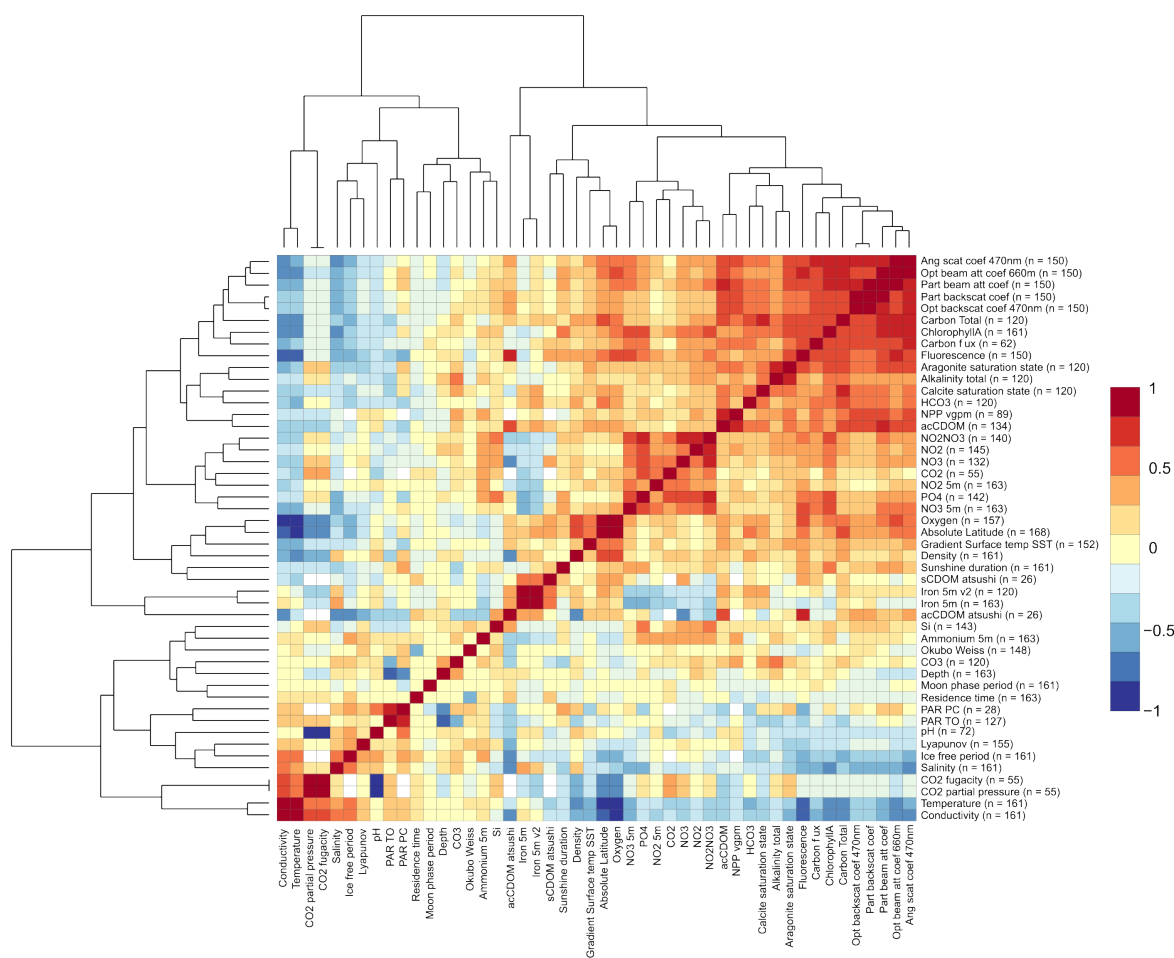


Figure S16: Correlation matrix for the environmental variables analyzed in the current work. Spearman rho correlation values are displayed. Sample size is indicated for each variable.