The genomic potential of photosynthesis in piconanoplankton is functionally redundant but taxonomically structured at a global scale

Alexandre Schickele, Pavla Debeljak, Sakina-Dorothée Ayata, Lucie Bittner, Eric Pelletier, Lionel Guidi, Jean-Olivier Irisson

ECOLOGY

The genomic potential of photosynthesis in piconanoplankton is functionally redundant but taxonomically structured at a global scale

Alexandre Schickele[1]*†, Pavla Debeljak[2,3], Sakina-Dorothée Ayata[4,5], Lucie Bittner[2,5], Eric Pelletier[6,7], Lionel Guidi[1,7]‡, Jean-Olivier Irisson[1,7]‡

Carbon fixation is a key metabolic function shaping marine life, but the underlying taxonomic and functional diversity involved is only partially understood. Using metagenomic resources targeted at marine piconanoplankton, we provide a reproducible machine learning framework to derive the potential biogeography of genomic functions through the multi-output regression of gene read counts on environmental climatologies. Leveraging the Marine Atlas of Tara Oceans Unigenes, we investigate the genomic potential of primary production in the global ocean. The latter is performed by ribulose-1,5-bisphosphate carboxylase/oxygenase (RUBISCO) and is often associated with carbon concentration mechanisms in piconanoplankton, major marine unicellular photosynthetic organisms. We show that the genomic potential supporting C_4 enzymes and RUBISCO exhibits strong functional redundancy and important affinity toward tropical oligotrophic waters. This redundancy is taxonomically structured by the dominance of Mamiellophyceae and Prymnesiophyceae in mid and high latitudes. These findings enhance our understanding of the relationship between functional and taxonomic diversity of microorganisms and environmental drivers of key biogeochemical cycles.

INTRODUCTION

Marine carbon fixation is largely performed by the piconanoplankton, responsible for 30 to 50% of global primary production (*1*, *2*). Piconanoplankton encompasses the unicellular eukaryotic marine plankton from the lower nano- to pico-size fractions (0.8 to 5 μm; also referred to as nano- and picoeukaryotes), including small diatoms, dinoflagellates, or prymnesiophytes. We hereafter refer to as piconanoplankton, following the Tara Oceans size fractions (*sensu 3*). These organisms are among the most diverse and abundant in the sunlit layer of the world ocean (*3*–*5*). In nutrient-poor areas, such as the oligotrophic open ocean, they locally contribute up to 80% of the phytoplanktonic biomass (*6*).

Most of the photosynthetic production on Earth relies on the ribulose-1,5-bisphosphate carboxylase/oxygenase [RUBISCO; (*7*)]. This enzyme is also responsible for photorespiration (Fig. 1). The latter is an energetically costly and metabolically inefficient pathway that consumes O_2 to produce CO_2 (*8*). However, RUBISCO does not clearly discriminate between CO_2 and O_2. RUBISCO emerged ~2 billion years ago in a period characterized by low oxygen (*9*). Therefore, its carboxylase function is unexpectedly inefficient relative to its oxygenase function, when considering the contemporary CO_2:O_2 ratio (*10*). The affinity of the carboxylase function relative to the oxygenase function of RUBISCO is referred to as the specificity factor (Fig. 1) that is variable across the tree of life, including marine phytoplankton (*9*, *11*). To compensate for the relative inefficiency of the carboxylase function of RUBISCO, carbon fixation pathways evolved ~30 million years ago when atmospheric CO_2 levels were estimated under 200 parts per million (ppm) (*12*, *13*). This induced selective pressure toward higher carbon fixation efficiency and led to the emergence of RUBISCO of higher specificity factor and various carbon concentration mechanisms (CCMs; i.e., biophysical or biochemical mechanisms). The latter aims to compensate for the specificity factor of RUBISCO by concentrating CO_2 at its active site (*8*).

Among biochemical CCMs, C_4 enzymes independently evolved across a large variety of marine and terrestrial lineages (*8*, *14*). The C_4 cycle is performed through three acid decarboxylation types (Fig. 1), all leading to an increase of the CO_2:O_2 ratio at the active site of RUBISCO (*15*): the malate decarboxylase–nicotinamide adenine dinucleotide phosphate (MDC-NADP) type, the MDC–nicotinamide adenine dinucleotide (MDC-NAD) type, and the phosphoenolpyruvate carboxykinase (PEPCK) type. The common enzyme to all C_4 acid decarboxylation types is phosphoenolpyruvate carboxylase (PEPC), fixing CO_2 in the cytosol by producing oxaloacetate (Fig. 1). In the MDC-NADP type, oxaloacetate is transferred to the chloroplast and reduced to malate. The latter is then decarboxylated, producing CO_2 and pyruvate, which is converted back to phosphoenolpyruvate (Fig. 1, blue pathway). In the MDC-NAD type, oxaloacetate is transferred to the mitochondria and reduced to malate. The decarboxylation reaction transfers CO_2 to the chloroplast by producing pyruvate that is transferred back to the chloroplast to be converted to phosphoenolpyruvate (Fig. 1, orange pathway). Last, the PEPCK type directly converts the mitochondrial oxaloacetate to phosphoenolpyruvate (Fig. 1, black pathway). However, it partially performs the MDH-NAD reduction and MDC-NADP decarboxylation reactions to balance the adenosine 5′-triphosphate (ATP) and NADPH budget, leading to common reactions and enzymes

[1]Sorbonne Université, CNRS, Laboratoire d'Océanographie de Villefranche, LOV, F-06230 Villefranche-sur-Mer, France. [2]Sorbonne Université, Muséum National d'Histoire Naturelle, CNRS, EPHE, Université des Antilles, Institut de Systématique, Evolution, Biodiversité (ISYEB), F-75005, Paris, France. [3]SupBiotech, Villejuif, France. [4]Sorbonne Université, CNRS, IRD, MNHN, Laboratoire d'Océanographie et du Climat, Institut Pierre Simon Laplace, LOCEAN-IPSL, F-75005 Paris, France. [5]Institut Universitaire de France, Paris, France. [6]Metabolic Genomics, Genoscope, Institut de Biologie François Jacob, CEA, CNRS, Université d'Evry, Université Paris Saclay, 91000 Evry, France. [7]Research Federation for the Study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans GOSEE, Paris, France.
*Corresponding author. Email: alexandre.schickele@imev-mer.fr
†Present address: Environmental Physics, Institute of Biogeochemistry and Pollutant Dynamics, ETH Zürich, 8092 Zürich, Switzerland.
‡These authors contributed equally to this work.

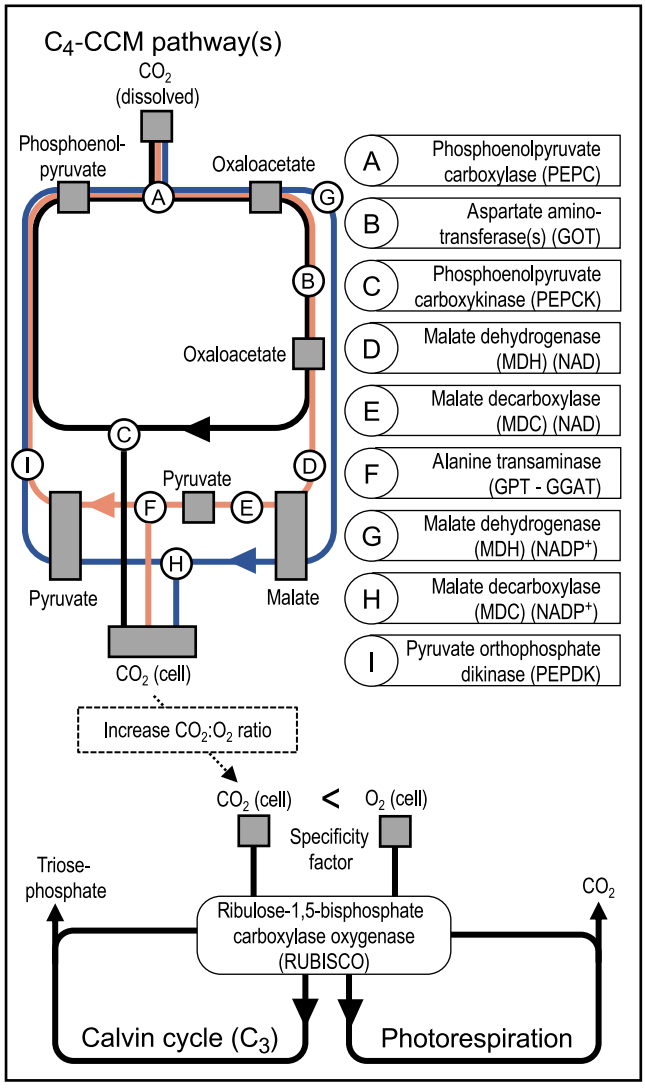

Fig. 1. Diagrammatic representation of the main enzymes and metabolites participating in the C_4 carbon concentration mechanisms, C_3 Calvin cycle, and photorespiration. Note that subcellular compartment and secondary metabolites are not represented. Enzyme names follow the Kyoto Encyclopedia of Genes and Genomes terminology. The three main currently described acid decarboxylation types are represented in blue [malate decarboxylase–nicotinamide adenine dinucleotide phosphate (MDC-NADP)], orange [MDC–nicotinamide adenine dinucleotide (MDC-NAD)], and black [phosphoenolpyruvate carboxykinase (PEPCK)], respectively.

between acid decarboxylation types (*15*). In the terrestrial realm, both physiological measurements and stable isotope techniques confirmed the presence of C_3 photosynthesis across a large range of environmental conditions, conversely to C_4 photosynthesis that is adapted to warm, nutrient-poor, and high irradiance conditions (*12*, *16*). In the marine realm, however, only a few studies explored the environmental affinity of C_4 photosynthesis regarding terrestrial-based hypotheses [see, e.g., (*13*, *14*, *17*)]. The potential for C_4 photosynthesis is highly suspected in key piconanoplankton lineages such as Mamiellophyceae and Prymnesiophyceae. Currently, subcellular evidence for C_4 enzymes include (i) MDC-NADP and PEPC in *Ostreococcus Tauri* (*18*); (ii) MDC-NADP, PEPC, three different oxoglutarate-to-malate translocator and pyruvate phosphate dikinase (PEPDK) in various *Micromonas* strains (*19*); and (iii) PEPC in the Prymnesiophyte *Emiliania huxleyi* [plastid presence and gene encoding (*20*)]. However, because of their small size (i.e., 0.8 to 5 μm) and poor representation in culture collections (*21*), physiological measurements and stable isotope analysis are lacking for natural piconanoplankton populations. Therefore, the genomic potential supporting C_3, and C_4 photosynthesis and its associated biogeography and functioning remains scarcely documented (*13*, *14*, *16*).

Recent global expeditions focusing on surface plankton sampling, together with advances in metagenomic sequencing, provided unique data to address the genomic potential and biogeography-related gaps [see, e.g., (*22*–*25*)]. In this context, metagenomics data are of growing interest to explore the hidden taxonomic and functional diversity potentially related to carbon fixation in piconanoplankton [see, e.g., (*26*, *27*)]. For example, genome-resolved metagenomics (*28*) based on the Tara Oceans eukaryotic metagenome led to the reconstruction of ~800 metagenome-assembled genomes [MAGs; (*29*)]. The latter are defined as genome-based taxonomic units, functionally and taxonomically annotated, and quantified by their associated genome-wide metagenomic reads. Therefore, MAGs offer the unique opportunity to study the genomic potential supporting carbon fixation and its biogeography, through both a functional and a taxonomic prism.

Habitat modeling is a popular niche theory–based tool to estimate species' biogeography according to the environmental conditions in which they are observed (*30*). Marine organisms are known for their important sensitivity to their surrounding environmental conditions, influencing growth, reproduction, and metabolic efficiency across all life stages (*31*). Thus, habitat modeling has been widely used to project the past, present, and future biogeography across various marine organisms, from zooplankton to fishes and marine mammals [see, e.g., (*32*)]. However, omics-based habitat modeling is still an emerging field to explore functional and taxonomic biogeography associated with unicellular planktonic organisms (*33*–*35*). Building on the abovementioned properties associated with MAGs, habitat modeling is transferable to genomic potential, thus exploring the quantitative response of the associated taxonomic and functional gene annotations to environmental conditions.

Here, complementing recent studies focusing on prokaryote or eukaryote-environment relationships (*26*, *33*, *34*), we provide an original, machine learning–based, comprehensive, and reproducible framework to derive the biogeography of the genomic potential related to metabolic functions, from metagenomic-based relative abundances data. Using multivariate boosted tree regressors [MBTRs; (*36*)], we simultaneously project the biogeography of selected genomic functional annotations while accounting both for their interactions and environmental responses. We applied this framework to metagenome-based protein functional clusters (PFCs; hereafter referred to as "clusters") linked to RUBISCO and C_4 enzymes only, in marine piconanoplankton. Compared to a more traditional approach (i.e., searching reads in a functional database using sequence similarity), our methodology combining MAGs and PFCs offers several advantages. The quantitative signal resulting from a MAG is (i) standardized by the genome length and (ii) corresponds to a taxonomic identity. Combined with PFCs, (iii) it also includes the fraction of signal corresponding to not yet annotated genes. Thus, this approach offers a more robust quantitative framework than

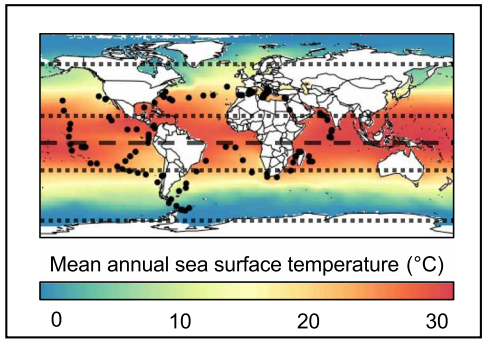

Fig. 2. Location of the Tara Oceans sampling stations. Stations are represented as black dots. The annual mean sea surface temperature from the World Ocean Atlas (*56*) is represented in the background. The dashed line corresponds to the equator. The dotted lines correspond to the 30°N and 60°N and 30°S and 60°S parallel, respectively.

traditional approaches, representative of eukaryotic plankton diversity in open oceans [39.1 billion reads recruited, ~97% identity, ~25 giga–base pair (Gbp) (*29*)] and transferable to a variety of functions or enzymes of interest using the already computed PFC network. Last, habitat modeling provides an interesting tool to estimate the response and co-dominance patterns of C_4 enzymes and RUBISCO to environmental conditions representative of the global ocean, conversely to estimates from the samples only, which might be driven by sampling and associated environmental biases.

RESULTS

C_4 CCM enzymes across sampled stations

From the Tara Oceans eukaryotic MAGs, ~1.2 million clusters were built, for which 349 are related to RUBISCO or C_4 enzymes within the 0.8- to 5-µm size fraction (fig. S1 and table S1). This dataset corresponds to 817 unique genes, with a median observed presence across 45 sampled stations per cluster. To avoid considering enzymes related to other metabolic functions, we selected those related to RUBISCO or C_4 enzymes only, corresponding to 240 clusters (fig. S1A), distributed across the world ocean; except the western Pacific and, to a lesser extent, Southern Ocean (Fig. 2). The successive cluster selection criteria (i.e., clusters exclusive to RUBISCO or C_4 enzymes, minimum presence at 10 sampling stations) did not present notable effects on the distribution of clusters across number of reads, number of genes, and taxonomic classes (fig. S2). In contrast, we observed a loss of signal for the MDCs (-NAD and -NADP) between functionally exclusive and nonexclusive clusters, highlighting an important fraction of sequence homologs for these enzymes (fig. S2).

Standardized distribution of the genomic potential related to C_4 photosynthesis

Here, we present projections for each C_4 enzyme and the RUBISCO. First, we rescaled the cluster-level projections (i.e., model outputs; fig. S1D) between 0 and 1 (i.e., distribution patterns; fig. S3). Then, we aggregated these patterns at the enzyme level according to their respective functional annotation. We therefore alleviated the propagation of the observed dominance of a given cluster to the aggregated enzyme-level patterns. The resulting enzyme-level projections are referred to as standardized patterns. For each enzyme, it represents a prediction of the genomic potential according to the environmental conditions at each geographical location and independently of any taxonomic dominance.

Because most C_4 enzymes are involved in several acid decarboxylation types, we cannot directly infer their corresponding distribution. However, MDC-NAD, MDC-NADP, and PEPCK are considered representative of their respective acid decarboxylation types. We predicted similar standardized patterns (Fig. 3) for all acid decarboxylation types and RUBISCO. The standardized patterns of all C_4 enzymes presented medium to high pairwise Pearson's correlation (0.5 to 0.9), except MDC-NAD and aspartate aminotransferase(s) (also called glutamic-oxaloacetic transaminase, GOT; Fig. 1) which are weakly correlated (0.3). We predicted a medium-to-high genomic potential (between 0.6 and 0.8) for most C_4 enzymes in temperate and tropical oligotrophic conditions (between 50°N and 40°S, excluding major upwelling areas; Fig. 3). The abovementioned predictions are associated with a coefficient of variation (CV) below 30% (Fig. 3A). The genomic potential of both PEPDK and MDC-NAD, however, presents lower values (between 0.3 and 0.4) in tropical oligotrophic gyres and in the pacific equatorial upwelling for PEPDK. Furthermore, we predicted areas of high genomic potential (>0.8) restricted to temperate areas such as the North and South Atlantic (~50°N and 40°S) and the North Pacific (~45°N) for RUBISCO and PEPCK, in comparison with other C_4 enzymes. These patterns suggest a higher affinity of the genomic potential of C_4 enzymes for the temperate and tropical oligotrophic conditions in comparison to RUBISCO. Furthermore, we predicted low-to-moderate potential (between 0 and 0.4) in high latitudes (i.e., above polar circles) for all standardized patterns (Fig. 3A). Predictions in such latitudes also present important calibration and projection-related variability, with coefficients of variations ranging from 30 to 100% (e.g., for the MDH-NADP and PEPCK). Therefore, our genomic potential predictions remain inconclusive in high latitudes, which are also subject to lower sampling coverage.

The environmental variables' importance in the trained model (fig. S4) highlighted the predominant roles of dissolved oxygen concentration (contributing to 34% of the explained variance) and the yearly variability (i.e., inter-month SD) in salinity (29%) and, to a lesser extent, of oxygen saturation, chlorophyll a concentration, and temperature. Furthermore, we revealed a strong affinity (i.e., maximum potential) of most standardized patterns (fig. S5) for tropical, oligotrophic conditions (e.g., temperature between 15° and 30°C; phosphate concentration below 0.5 µmol/kg). However, we predicted different responses to the variability in chlorophyll a concentration and euphotic zone depth across enzymes (fig. S5). Last, we highlighted no taxonomic dominance across the world oceans, according to the taxonomic composition associated with each cluster, suggesting a worldwide functional redundancy in the genomic potential supporting C_4 enzymes in piconanoplankton (fig. S6).

Weighted distribution of the genomic potential related to C_4 photosynthesis

Here, we present projections for each C_4 enzyme and the RUBISCO. First, we rescaled the cluster-level projections (i.e., model outputs; fig. S1D) by their observed metagenomic read abundance (i.e., weighted distribution patterns; fig. S3). Then, we aggregated these patterns at the enzyme level according to their respective functional annotation. We therefore propagate the observed dominance of a

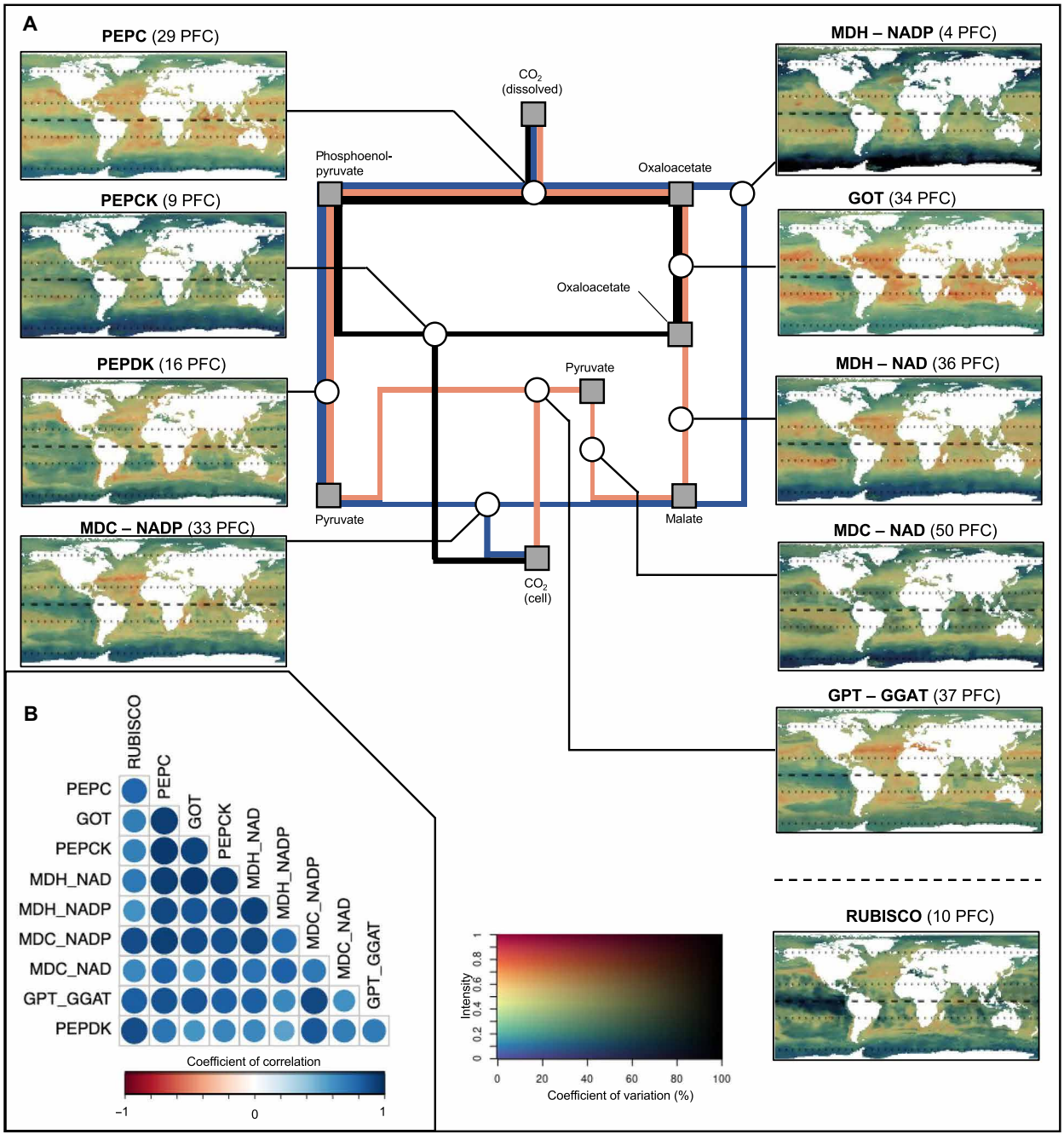

Fig. 3. Standardized distributions of the genomic potential. Standardized patterns corresponding to the relative genomic potential supporting C_4 enzymes and RUBISCO. (**A**) Synthetic diagram of the metabolic pathway and corresponding projections. (**B**) Inter-projections Pearson's spatial correlation index. The three main currently described acid decarboxylation types are represented in blue (MDC-NADP), orange (MDC-NAD), and black (PEPCK), respectively. Involved metabolic components and enzymes are indicated on the diagram by squares and circles, respectively. The two-dimensional (2D) color scale represents the standardized genomic potential for the target enzyme as the hue value (*y* axis) and the associated coefficient of variation as the saturation (i.e., uncertainty in % of the mean; *x* axis). An orange-to-red hue corresponds to a region where environmental conditions yield a high proportion (>0.6) of the target genes in the model. A low saturation level corresponds to an important variance among the underlying cluster-level projections. The dashed line on the projections corresponds to the equator. The dotted lines on the projections correspond to the 30°N and 60°N and 30°S and 60°S parallel, respectively.

given cluster (i.e., and associated taxa) to the aggregated enzyme-level patterns. The resulting enzyme-level projections are referred to as weighted patterns. For each enzyme, it represents the corresponding genomic potential (i.e., relative to the other considered enzymes), according to the environmental conditions at each geographical location.

We predicted contrasting weighted patterns between the RUBISCO and the acid decarboxylation types (Fig. 4A). The weighted pattern of RUBISCO presented maximum potential in temperate areas (Fig. 4B). We predicted low-to-moderate potential (<0.3) and moderate (~30%) uncertainty in high latitudes for the weighted patterns of PEPC, MDCs, MDHs, and transferases (i.e., GOT and alanine

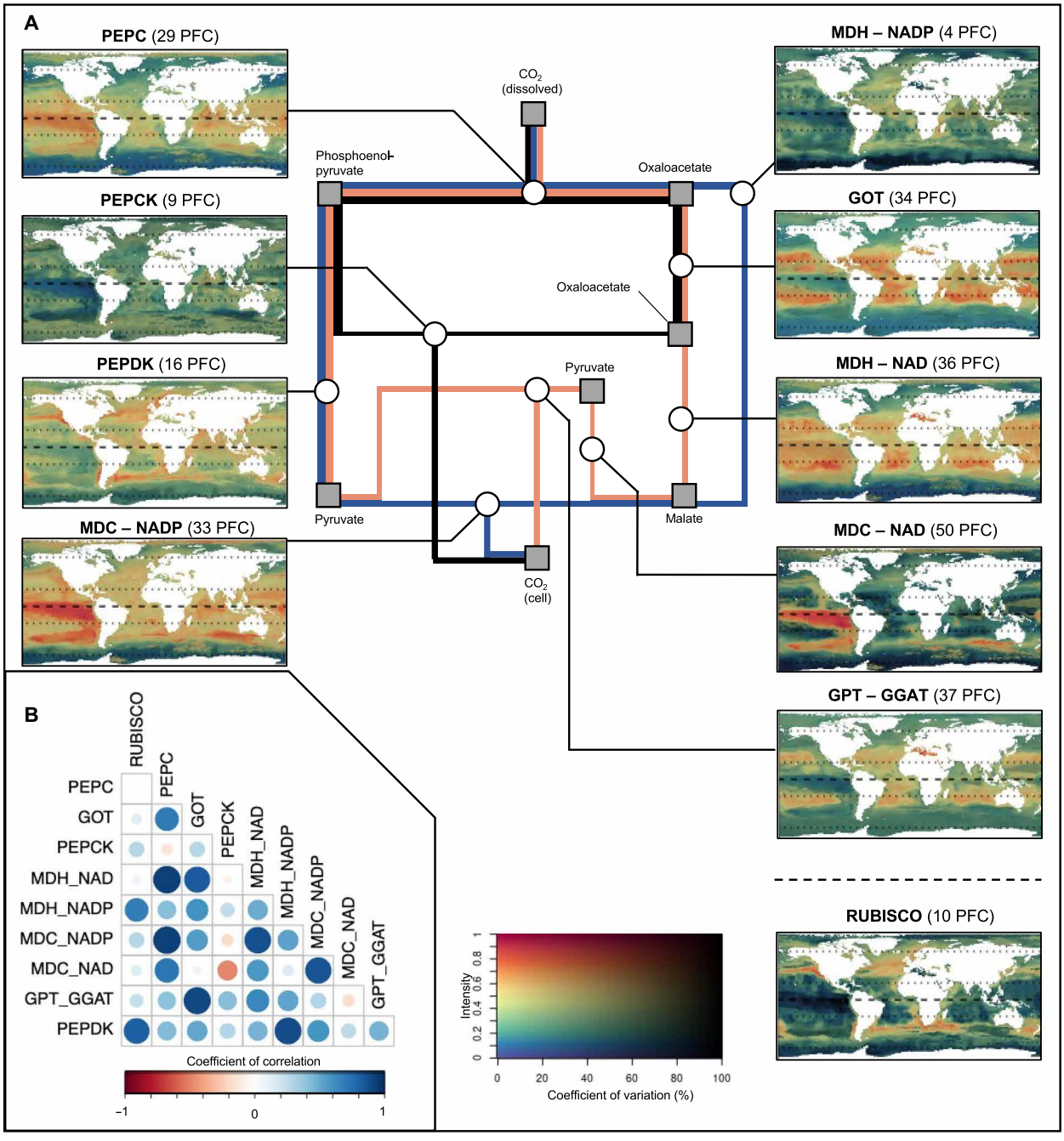

Fig. 4. Weighted distributions of the genomic potential. Weighted patterns corresponding to the relative genomic potential supporting C_4 enzymes and RUBISCO, rescaled by the corresponding observed relative metagenomic reads abundance. (**A**) Synthetic diagram of the metabolic pathway and corresponding projections. (**B**) Inter-projections Pearson's spatial correlation index. The three main currently described acid decarboxylation types are represented in blue (MDC-NADP), orange (MDC-NAD), and black (PEPCK), respectively. Involved metabolic components and enzymes are indicated on the diagram by squares and circles, respectively. The 2D color scale represents the weighted genomic potential for the target enzyme as the hue value (*y* axis) and the associated coefficient of variation as the saturation (i.e., uncertainty in % of the mean; *x* axis). An orange-to-red hue corresponds to a region where environmental conditions yield a high proportion (>0.6) of the target genes in the model. A low saturation level corresponds to an important variance among the underlying cluster-level projections. The dashed line on the projections corresponds to the equator. The dotted lines on the projections correspond to the 30°N and 60°N and 30°S and 60°S parallel, respectively.

transaminase; GPT-GGAT; Fig. 4A). These patterns also presented moderate-to-high potential (between 0.5 and 1) in tropical areas, with some discrepancies. We show a Pearson's correlation index above 0.5 between the abovementioned enzymes and above 0.7 for GOT and MDHs (Fig. 4B). The latter presented an important potential in oligotrophic regions (e.g., Pacific gyres), suggesting functional redundancy in the genomic potential from oxaloacetate to malate (Fig. 4A). In contrast, we predicted a high potential (>0.7) in eutrophic Pacific waters for the weighted patterns of MDCs (Pearson's correlation above 0.7; Fig. 4A). Overall, we show high confidence in the areas associated with high genomic potential, with CVs lower than 30% among all trained algorithms and 100 bootstrap projections. The abovementioned weighted responses to environmental variables are like the ones highlighted in the previous section, characterized by higher potential in warm, low seasonality, and generally oligotrophic water bodies (figs. S7 and S8).

Conversely, we predicted moderate to high-intensity values in oligotrophic tropical areas, but in the Southern Ocean (>0.5; Fig. 4) for the weighted pattern of PEPCK (i.e., a different acid decarboxylation type). The latter was preferentially distributed along water bodies characterized by (i) high seasonality of the chlorophyll a concentration and the depth of the euphotic zone, (ii) high concentrations of oxygen (presenting the highest explanatory power in the model training; fig. S4) and nutrients (e.g., phosphates and nitrates), and (iii) average temperatures below 8°C (fig. S7).

Last, weighted patterns associated with high latitudes (e.g., correlated with the one of PEPCK) were composed at 28% of Prymnesiophyceae and 50% of Mamiellophyceae (Shannon index of 1.5), based on the taxonomic composition of each cluster. Mamiellophyceae also composed 40% of the patterns with a clear temperate affinity (e.g., correlated with the one of RUBISCO; fig. S8). In contrast, a larger diversity of taxonomic classes, with a Shannon index of 2.1, was obtained for patterns associated with equatorial latitudes.

DISCUSSION

Genomic potential for C_4 CCM in piconanoplankton

By selecting clusters annotated by C_4 enzymes or RUBISCO only, we considered a fraction of the available metagenomic information (i.e., ~67% of all the clusters related to C_4 enzymes or RUBISCO). In addition, genes related to other metabolic pathways may have responses to environmental variables different from genes related to C_4 enzymes, potentially including bias in their corresponding clusters' projection. Therefore, selecting a reduced set of clusters alleviates the risk of metabolic noise in the environmental responses, limited to the effect of C_4 enzymes potentially involved in other pathways (e.g., GPT-GGAT transporter).

Our study focused on piconanoplankton, the photosynthetic fraction of which is generally dominated by the Prymnesiophyceae, Bacillariophyceae, Dinophyceae, and Mamiellophyceae lineages in the open ocean (*3*, *21*). The latter is a major clade of the polyphyletic Prasinophyceae assemblage (*37*). The potential for C_4 photosynthesis has been suggested for several families, including Bacillariophyceae by combining C_4 enzyme inhibition and photosynthetic efficiency monitoring [e.g., PEPDK (*38*), PEPC, and PEPCK (*39*)]. Evidence for genes encoding all C_4 enzymes exists in *Micromonas* and *Ostreococcus*, both belonging to the Mamiellophyceae (*37*, *40*). A plastid PEPC enzyme was recently found in *E. huxleyi* (*38*), a Prymnesiophyte abundant in temperate and polar regions (*41*). However, to our knowledge, no study provided univocal evidence for C_4 CCM usage in natural populations for the smallest fraction of piconanoplankton, contrasting with recent findings supporting C_4 CCM usage by marine diatoms (*14*). Stable isotope measurements would be necessary to fully understand C_4 photosynthesis in piconanoplankton, but they are difficult to apply at the species level in natural, uncultured, plankton communities [see, e.g., (*16*, *17*)]. Alternatively, recent literature suggests the need for further studies on deep chlorophyll a maxima and various transporters (e.g., bicarbonate transporters), some of which are associated with or specific to C_4 metabolism, to better understand C_4 CCM in natural populations (*14*, *15*).

Complementing these experimental approaches, we use a data-driven approach to shed more light on the environmental drivers of C_4 genes in marine piconanoplankton. However, MAGs integrate chloroplast and mitochondrial genes corresponding to C_4 enzymes but do not distinguish their origin (*29*) nor provide information on the subcellular location of the corresponding enzymes (*13*, *42*). Therefore, the patterns presented here must be interpreted as the potential for the (co-)presence of those pathways in the genome. They should now be complemented by culture-based studies, locating enzymes within cells and/or performing carbon isotope discrimination to confirm C_4 CCM presence, expression, and its coexistence with C_3 photosynthesis in piconanoplankton lineages (*16*). The present study could be used to locate regions where such mechanisms are most likely to occur.

Environment-driven genomic potential

The modeled distribution patterns revealed that the genomic potential for C_4 photosynthesis is more associated with tropical oligotrophic and annually stratified waters. Conversely, the proportion of reads related to RUBISCO (i.e., considered a representative of all photosynthetic pathways, due to its central role in C_3, C_4, and CAM photosynthesis) is higher in temperate regions (Fig. 3A). The fact that terrestrial C_4 plants (*8*) and the genomic potential for C_4 CCM in piconanoplankton display similar latitudinal distribution, around the tropics, does not imply that the environmental drivers of those distributions are the same. In terrestrial plants, C_4 CCMs are considered an adaptation to drought and are, for example, also associated with a specific leaf structure that reduces their water consumption (*8*). Drought is of course not an evolutionary driver for marine piconanoplankton. Alternatively, they present an important surface area:volume ratio [i.e., small cells or presence of a vacuole (*43*, *44*)] leading to a high nutrient absorption yield, which is adapted to oligotrophic waters, common in the tropical ocean.

In addition to environmental conditions, the biogeography of the genomic potential supporting C_4 CCM may also relate to irradiance levels, largely controlling ATP generation, necessary for the decarboxylation reaction (*43*). C_4 CCM requires additional ATP generation to increase the RUBISCO efficiency in comparison to classical C_3 photosynthesis without affecting the energy available for the latter (*43*, *45*). In contrast, an excess of ATP may lead to photoinhibition, thus lower carbon fixation efficiency (*38*, *46*). Therefore, it has been suggested that C_4 photosynthesis is particularly adapted to dissipate excess energy in the cell in high irradiance areas such as tropical oceans (*14*, *38*). Our weighted patterns highlighted differences between PEPCK and MDCs (Fig. 4). The latter requires two extra ATPs compared to the C_3 carbon fixation to complete the pathway. In a logical way, the PEPCK acid decarboxylation type, which only requires one extra ATP and thus is supposed to be more efficient in low irradiance

environments (*45*), showed here the highest genomic potential in polar or subpolar regions.

Functional and ecological implications

We highlighted functional redundancy among C_4 genes in oligotrophic tropical waters (fig. S6). This contrasts with high latitudes, where only a few taxa dominate (fig. S8) (*4*, *47*). We highlighted a biogeographical differentiation between the weighted pattern of RUBISCO—i.e., the baseline photosynthetic enzyme—and those of C_4 enzymes. In the period ranging from 30 million to 20,000 years ago, the average atmospheric CO_2 concentration markedly reduced from ~1000 ppm to less than 200 ppm. This long-term atmospheric CO_2 concentration trend induced a lower concentration of dissolved inorganic carbon in the surface ocean waters (*8*). This led to a selective pressure toward efficient photosynthetic metabolism, like C_4 CCMs (*12*) or, to a lesser extent, RUBISCO of higher carboxylation affinity [e.g., type II in Dinoflagellates (*13*)]. While the evolution of C_4 CCMs in marine organisms is not yet fully understood, 48 independent evolutions of C_4 CCMs were identified in the genome of terrestrial plants [e.g., grasses and caryophyllales (*8*)], suggesting a higher genomic potential for C_4 CCMs in taxonomically diverse areas (*12*). The abovementioned functional redundancy in the genomic potential for C_4 CCM in taxonomically rich tropical waters may relate to coevolution between taxonomic diversification and its associated functions (i.e., neutral theory). However, the functional diversity among C_4 acid decarboxylation types may also reflect—or be amplified by—a selection process, as it may present a selective advantage. Moreover, the dominance of Mamiellophyceae (i.e., relative to other piconanoplankton associated with RUBISCO or C_4 enzymes) in the temperate and polar latitudes (associated with the patterns of RUBISCO and PEPCK; fig. S8) is concordant with the literature (*48*). Although cosmopolitan, several Mamiellophyceae species have been shown as important in both the Arctic and Antarctic [see, e.g., (*49*–*51*)]. Likewise, the cosmopolitan Prymnesiophyceae has been identified as dominant in high latitudes (associated with the pattern of PEPCK; fig. S8), including in the Southern Ocean (*41*) and associated with Mamiellophyceae (i.e., Prasinophyceae) in the North Atlantic (*52*, *53*). The literature therefore validates their predicted biogeography. We identified key environmental predictors shaping the biogeography and (co-)dominance patterns of the genomic potential supporting C_4 enzymes and RUBISCO in marine piconanoplankton. Such results open perspectives for exploring the relationship between functional and taxonomic diversity in the oceans, complementing already diverse approaches and data types, and for a better understanding of the environmental drivers of key biogeochemical cycles in the current and future climatic context.

MATERIALS AND METHODS

Experimental design

Genomic and environmental data

We studied the biogeography of the genomic potential related to C_4 enzymes through the prism of MAGs (*29*) retrieved from the Tara Oceans expedition (2009–2013). Briefly, 280 billion reads from 798 metagenomes, corresponding to the surface and deep chlorophyll maximum layer of 210 stations from the Pacific, Atlantic, Indian, Southern, and Arctic Oceans, as well as the Mediterranean and Red Seas (Fig. 2), encompassing eukaryote-enriched plankton size fractions ranging from 0.8 μm to 2 mm, were used as inputs for 11 metagenomic coassemblies (6 to 38 billion reads per coassembly) using geographically bounded samples. We thus created a culture-independent, nonredundant (average nucleotide identity <98%) genomic database for eukaryotic plankton in the sunlit ocean consisting of 683 MAGs and 30 single-cell genomes, all containing more than 10 million nucleotides for a total size of 25.2 Gbp and encoding for 10,207,450 genes. Then, a sequence similarity network (SSN) was built out using the 683 manually curated MAGs following a similar methodology to the one developed in (*33*). A pairwise comparison was computed between each protein sequence. The resulting alignment was then filtered, removing self-hits and pairs showing less than 80% of sequence identity and coverage. Resulting PFCs [as in (*33*)] were built, hereafter referred to as clusters. A functional annotation was added to the sequences, and the functional homogeneity was checked in each cluster (*54*, *55*).

For each of the 130 selected Tara Oceans metagenomic surface samples, we retrieved a set of monthly, global scale, environmental climatologies (*56*–*58*) encompassing the 2005 to 2017 period, at a spatial resolution of 1° × 1° (table S2). The latter corresponds to the available climatology encompassing the sampling period (2009–2013), where we considered temporal environmental variations negligible in comparison to spatial environmental gradients. They correspond to a restricted set of factors characterizing the water body (e.g., oligotrophic and eutrophic) and related to C_4 photosynthesis, for which we calculated the yearly average and yearly SD (i.e., a proxy of seasonal variations).

Protein functional cluster selection and preprocessing

We first selected a reduced set of clusters, within the 0.8- to 5-μm size fraction, for which 100% of the Kyoto Encyclopedia of Genes and Genomes orthology (*59*) annotated protein members were related to C_4 enzymes or RUBISCO (fig. S1 and table S1). To avoid model overparameterization and because rare clusters were assumed as not influencing the large-scale patterns investigated in this study, we only considered clusters that were present in a minimum of 10 Tara Oceans stations. The corresponding dataset contained 240 clusters, associated with 234 MAGs. The latter presented an average completeness estimate of 57% (data S1). In comparison, the average completeness estimate across all MAGs from Delmont *et al.* (*29*) yields 37%. As a supplementary quality check, we estimated a minimum horizontal coverage (i.e., the number of bases of a MAG covered with a certain depth) of 68% for each of the 234 MAGs (data S1). Last, we assessed the quality of our MAGs using the Benchmarking Universal Single-Copy Orthologs (BUSCO) protocol (*60*). The latter is a set of conserved single-copy genes present in most eukaryotic and prokaryotic genomes. It is used to assess both the completeness and quality of genomic data by comparing the presence of conserved single-copy genes across genomes (i.e., the percentage of mapped BUSCO genes in each MAG). It therefore complements technical metrics such as contiguity measures and is largely applicable across datasets (*60*). We show that our MAGs are associated with an average BUSCO completeness of 55.7% (data S1). We therefore consider these MAGs of sufficient quality for identifying C_4 genes across our samples. To reduce the number of response variables (clusters; PFCs) to a reasonable amount for multivariate modeling, with respect to the limited number of stations, we performed an Escoufier dimensional reduction (*61*). The latter iteratively selects the clusters whose pattern across stations minimizes the residual variance of the dataset. Here, we selected 50 clusters that represent more than 95% of the 240 clusters variance to be

included in the multivariate algorithm. To alleviate the effect of gene length and sequencing effort variability between samples on the number of reads, we normalized the metagenomic reads by the length of the corresponding gene coding part and the total number of reads per station (i.e., including reads of all non-considered clusters), respectively. Because the total genomic material present at each sampling station is unknown (i.e., non-exhaustive sampling and sequencing effort), the absolute number of reads is not comparable among stations. To compare the abundance between selected clusters at different sampling stations, we transformed the dataset to relative abundance (fig. S1).

Multivariate boosted tree regressors

General principle. Recently, growing interest in interactions between response variables led to the development of multivariate machine learning algorithms, such as MBTRs (*36*). The latter is particularly adapted to a small sample size as the interactions between response variables are considered supplementary information to calibrate the model. Here, we use MBTR to model the relationship between climatologies and metagenomic relative abundance (i.e., summed at 1 for each station; fig. S1). To best reproduce the response of metagenomic reads (i.e., response variable) to the corresponding environmental variables (i.e., explanatory variable), the model sequentially fits decision trees (i.e., boosting rounds) using gradient descent to minimize a specific loss function. At each boosting round, the algorithm fits a decision tree on the residuals of the previous boosting round and computes a tree loss (i.e., a measure of deviation between observed and predicted response variable values). Decision trees are constructed using the hessian of the loss function (i.e., second-order tensor of its partial derivatives) to minimize the loss gradient. Therefore, the information learned by the nth tree is passed to the n+1th tree at a user-defined learning rate (fig. S1). The ensemble of sequentially fitted decision trees is considered in the model until the minimum loss is reached. Last, one important feature of MBTR is the conservation of the initial correlation structure between the response variables [see methods in (*36*)].

Model training and evaluation. To avoid overfitting, the explanatory and response datasets were split between the training set and the test set using a n-fold cross-validation procedure. For each model, n algorithms were trained on different n−1-folds, while the remaining fold was used for testing only (i.e., computing the loss at each boosting round). To minimize the effect of spatial and temporal autocorrelation in our data [i.e., leading to overoptimistic model evaluation (*62*)] the n-folds were defined according to the Tara Oceans station number. The latter follows a continuous trajectory in space and time, resulting in spatially and temporally distant folds [i.e., spatial and temporal block splitting, as recommended in (*62*)]. The resulting n algorithms predictions were aggregated in an average response and its corresponding CV. The ability of the final model to reproduce the observed clusters' relative abundance across environmental conditions has been measured by the R^2 criteria and the root mean square error (RMSE; between 0 and 1 according to the distribution pattern scale).

Spatial projections. To better estimate projection uncertainty, our spatial projections were constructed using a bootstrap procedure. For each 100-bootstrap round, we first resampled the original dataset (i.e., train and test response dataset and corresponding explanatory variable values) with replacement. Then, we refitted an MBTR algorithm on the resampled data by using the hyperparameters corresponding to the validated model, including the number of boosting rounds corresponding to the minimum loss across all n algorithms. Last, the refitted MBTR algorithm was used to predict the relative abundance of clusters worldwide, using the corresponding climatological values at each geographical cell.

From model projections to final outputs

We only modeled the 50 clusters representing 95% of the dataset variability. Therefore, we indirectly reconstructed the projections of the 190 others by identifying their most representative Escoufier-selected cluster. To this extent, we performed a correspondence analysis based on the observed relative abundance of all clusters. By using the dimensions of the correspondence analysis space corresponding to a minimum of 80% variance explained, we calculated the Euclidean distance between each nonselected cluster, and its nearest neighbor selected by the Escoufier criteria. Because the 50 Escoufier-selected clusters represented more than 95% of the dataset variability, we considered that a cluster and its nearest neighbor in the correspondence analysis space share the same relative abundance pattern. We then reconstructed the spatial projections of the 190 clusters not considered in MBTR according to their projected nearest Escoufier-selected neighbor. The resulting 240 cluster-level projections of the genomic potential were then aggregated at the enzyme level according to their functional annotation (see Results; fig. S3). Each projection was standardized between 0 and 1, thus considered equally weighted distribution patterns (fig. S3, top). We then performed two aggregation methods, leading to a (i) standardized and a (ii) weighted distribution of the genomic potential related to C_4 photosynthesis. In the former, we performed a simple average of all cluster-level projections sharing a similar functional annotation (fig. S3, left). This resulted in an enzyme-level projection reflecting the most common patterns at the cluster level. In other words, the highest genomic potential at the enzyme level was located where most cluster-level projections present their highest genomic potential (fig. S3, left). It was independent of any taxonomic dominance. In the latter, however, we performed a weighted average of all cluster-level projections sharing a similar functional annotation. The weights corresponded to the sum of the observed relative abundance of each cluster, across all stations (figs. S1A and S3, right). This resulted in an enzyme-level projection reflecting the cluster-level patterns with the highest relative abundance (i.e., dominant patterns). In other words, the highest genomic potential at the enzyme level was located where abundant cluster-level projections presented their highest genomic potential (fig. S3, left). It propagated the associated taxonomic dominance to the enzyme-level patterns.

Statistical analyses

Metagenomic data construction

The bioinformatic workflow designed to build the MAGs can be found in (*63*) and on the genoscope website: www.genoscope.cns.fr/tara/. Original metagenomes are available under the European Bioinformatics Institute repository with project ID PRJEB402 and organized into four major size classes of 0.8 to 5 μm, 3 to 20 μm, 20 to 180 μm, and 180 to 2000 μm. To estimate the abundance and expression of each contig in each sample, cleaned reads (from metagenomes and metatranscriptomes) were mapped against the eukaryotic MAGs using the bwa tool (version 0.7.4). The following parameters were used: bwa aln -l 30 -O 11 -R 1; bwa sample -a 20000 -n 1 –N; samtools; rmdup. Reads covering at least 80% of read length with at least 95% of identity were retained for further analysis. In the case of several possible best matches, a random one was picked. The first SSN was built out of 683 manually curated

MAGs from 10,207,435 eukaryotic proteins. This file was used for the creation of a diamond database and a protein blast of the protein sequences against the database to compute the percentage of similarity between every pair of proteins detected in the MAGs. Here, we used a maximum *e*-value of 1×10^{-3} and the sensitive option adapted to long reads (−e $1e^{-3}$ −p 30–sensitive). The alignment was then filtered removing all self-hits. Several thresholds for the percentage of identity and coverage were tested (75, 80, 85, and 90%). An SSN (bioinformatic workflow available at https://data.d4science.net/BN9t) was built with the diamond output using 80% identity and 80% coverage threshold to minimize the number of singletons while maximizing the functional homogeneity between linked proteins. Reproducible analysis and statistical exploration are provided in (*33*, *54*, *64*). An SSN is made of singletons (vertices or sequences without any homology with other sequences) and connected components (CCs; i.e., subgraphs composed of at least two vertices disconnected from the rest of the network). In our case, a CC corresponds to a group of at least two protein sequences that are linked together (directly or via neighbors) and that have no link with other groups of sequences in the SSN. We assume that the proteins contained in a CC potentially share a similar molecular function (*54*, *55*, *64*, *65*). These proteins were functionally annotated using eggNOG mapper v2.1.5.

Data selection

Unless specified, all following analyses were performed using R 3.14, with the corresponding code and libraries available at https://data.d4science.net/qa7Z or 10.5281/zenodo.11093527. Starting from station no. 66 (i.e., Cape Town), a supplementary size class of 0.8 to 2000 µm has been implemented in the Tara Oceans cruise, while the initial 0.8 to 5 µm was not sampled from stations 155 onward (i.e., Arctic stations). Given that smaller organisms are much more abundant than large ones, the majority of organisms sampled with the 0.8- to 2000-µm filter are piconanoplankton (i.e., corresponding to 0.8 to 5 µm). We tested this hypothesis by analyzing the composition, detection rate, and associated percentage of metagenomic reads across clusters associated with RUBISCO or C_4 enzymes, between 0.8 to 5 µm and 0.8 to 2000 µm at their common sampling stations (i.e., 66 to 155). The composition of clusters of interest across common stations presented a significant correlation (Pearson correlation: 0.89; *P* value: 0.01) between both size fractions. Moreover, 85.5% of the abovementioned clusters detected in the 0.8- to 2000-µm fraction were also detected in the 0.8- to 5-µm fraction. The latter represents 86.3% of the mapped reads in the 0.8- to 2000-µm fraction. The clusters detected in both fractions at common locations present a Pearson and Spearman metagenomic read correlation of 94.4 and 87.3%, respectively. This supports the inclusion of Arctic data issued from the 0.8- to 2000-µm filter. The effect of the different selection criteria such as the exclusivity and the minimum number of stations coverage is shown in fig. S3 and calculated using Pearson's chi-square test for count data ("chisq.test" function; $P < 0.05$).

Model training, evaluation, and projections

We fitted one MBTR (Python >3.7.) algorithm per training set and hyperparameter combination, under a mean square error loss function, a learning rate of 5.10^{-3}, and a 10-fold cross-validation procedure. We set the minimum number of observations in terminal leaves to 30 and the number of quantiles considered to find the best split to 10. The model loss is adapted to discriminate between different sets of hyperparameters. Thus, we only considered the set of hyperparameters that resulted in the minimum loss across the corresponding 10 trained MBTR algorithms. However, the loss does not provide information on the actual performance of the model in reproducing the observed data. Therefore, for each of the 10 trained MBTR algorithms, we predicted the relative abundance of the metagenomic reads on the environmental values corresponding to each of the 10 test sets. The 10 corresponding predictions were compared against the truth (i.e., observed values) by means of the *R*-squared (R^2) and RMSE (here between 0 and 1) to assess model performance on data not seen by the MBTR models during the training process. The corresponding model evaluation estimated an R^2 of 0.33 and an RMSE of 0.05. The conservation of the correlation structure in MBTR was tested by computing a Pearson correlation matrix between response variables before and after model fitting, followed by a Mantel test (Pearson's $R = 0.748$; $P = 0.01$). Last, for the spatial projections, we performed a total of 100 bootstrap rounds and computed the average and CV between all bootstrap projections (fig. S1).

Model outputs

The variable importance in the model training (fig. S4) was calculated as the number of times an environmental variable was selected for a tree split, scaled by the corresponding loss gain. The response of the genomic potential for an enzyme to each environmental variable was estimated by partial dependence plots. The latter was defined as the marginal response of the target to a feature over the values of all other input features. To estimate the taxonomic composition associated with each pattern (i.e., standardized, or weighted), we constructed the distribution pattern of each MAG using the same modeling framework. We then performed a hierarchical clustering [the "ward.D2" method (*66*)] on MAGs level projections that were then correlated to the enzyme's distributional patterns (Pearson's *R* correlation). Last, according to each MAG annotation, we computed the taxonomic composition corresponding to each cluster of MAG patterns.

Supplementary Materials

This PDF file includes:

Figs. S1 to S8
Tables S1 and S2
Legend for data S1

Other Supplementary Material for this manuscript includes the following:

Data S1

REFERENCES AND NOTES

1. E. Granum, J. A. Raven, R. C. Leegood, How do marine diatoms fix 10 billion tonnes of inorganic carbon per year? *Can. J. Bot.* **83**, 898–908 (2005).
2. A. Z. Worden, F. Not, Ecology and diversity of picoeukaryotes, in *Microbial Ecology of the Oceans*, D. L. Kirchman, Ed. (John Wiley & Sons, Inc., 2008), pp. 159–205.
3. C. de Vargas, S. Audic, N. Henry, J. Decelle, F. Mahé, R. Logares, E. Lara, C. Berney, N. Le Bescot, I. Probert, M. Carmichael, J. Poulain, S. Romac, S. Colin, J.-M. Aury, L. Bittner, S. Chaffron, M. Dunthorn, S. Engelen, O. Flegontova, L. Guidi, A. Horák, O. Jaillon, G. Lima-Mendez, J. Lukeš, S. Malviya, R. Morard, M. Mulot, E. Scalco, R. Siano, F. Vincent, A. Zingone, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, T. O. Coordinators, S. G. Acinas, P. Bork, C. Bowler, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, F. Not, H. Ogata, S. Pesant, J. Raes, M. E. Sieracki, S. Speich, L. Stemmann, S. Sunagawa, J. Weissenbach, P. Wincker, E. Karsenti, Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).
4. F. M. Ibarbalz, N. Henry, M. C. Brandão, S. Martini, G. Busseni, H. Byrne, L. P. Coelho, H. Endo, J. M. Gasol, A. C. Gregory, F. Mahé, J. Rigonato, M. Royo-Llonch, G. Salazar, I. Sanz-Sáez, E. Scalco, D. Soviadan, A. A. Zayed, A. Zingone, K. Labadie, J. Ferland, C. Marec, S. Kandels, M. Picheral, C. Dimier, J. Poulain, S. Pisarev, M. Carmichael, S. Pesant, S. G. Acinas, M. Babin, P. Bork, E. Boss, C. Bowler, G. Cochrane, C. de Vargas, M. Follows, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels, L. Karp-Boss, E. Karsenti, F. Not, H. Ogata, S. Pesant, N. Poulton, J. Raes, C. Sardet, S. Speich,

L. Stemmann, M. B. Sullivan, S. Sunagawa, P. Wincker, M. Babin, E. Boss, D. Iudicone, O. Jaillon, S. G. Acinas, H. Ogata, E. Pelletier, L. Stemmann, M. B. Sullivan, S. Sunagawa, L. Bopp, C. de Vargas, L. Karp-Boss, P. Wincker, F. Lombard, C. Bowler, L. Zinger, Global trends in marine plankton diversity across kingdoms of life. *Cell* **179**, 1084–1097.e21 (2019).
5. A. Obiol, C. R. Giner, P. Sánchez, C. M. Duarte, S. G. Acinas, R. Massana, A metagenomic assessment of microbial eukaryotic diversity in the global ocean. *Mol. Ecol. Resour.* **20**, 718–731 (2020).
6. R. Massana, Eukaryotic picoplankton in surface oceans. *Annu. Rev. Microbiol.* **65**, 91–110 (2011).
7. Y. M. Bar-On, R. Milo, The global mass and average rate of rubisco. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 4738–4743 (2019).
8. R. F. Sage, T. L. Sage, F. Kocacinar, Photorespiration and the evolution of C_4 photosynthesis. *Annu. Rev. Plant Biol.* **63**, 19–47 (2012).
9. P. M. Shih, A. Occhialini, J. C. Cameron, P. J. Andralojc, M. A. J. Parry, C. A. Kerfeld, Biochemical characterization of predicted precambrian RuBisCO. *Nat. Commun.* **7**, 10382 (2016).
10. T. J. Erb, J. Zarzycki, A short history of RubisCO: The rise and fall (?) of nature's predominant CO_2 fixing enzyme. *Curr. Opin. Biotechnol.* **49**, 100–107 (2018).
11. R. P. Haslam, A. J. Keys, P. J. Andralojc, P. J. Madgwick, A. Inger, A. Grimsrud, H. C. Eilertsen, M. A. J. Parry, Specificity of diatom Rubisco, in *Plant Responses to Air Pollution and Global Change*, K. Omasa, I. Nouchi, L. J. De Kok, Eds. (Springer, 2005), pp. 157–164.
12. R. F. Sage, M. Stata, Photosynthetic diversity meets biodiversity: The C_4 plant example. *J. Plant Physiol.* **172**, 104–119 (2015).
13. J. R. Reinfelder, Carbon concentrating mechanisms in eukaryotic marine phytoplankton. *Ann. Rev. Mar. Sci.* **3**, 291–315 (2011).
14. J. J. Pierella Karlusich, C. Bowler, H. Biswas, Carbon dioxide concentration mechanisms in natural populations of marine diatoms: Insights from Tara Oceans. *Front. Plant Sci.* **12**, 657821 (2021).
15. R. T. Furbank, Evolution of the C_4 photosynthetic mechanism: Are there really three C_4 acid decarboxylation types? *J. Exp. Bot.* **62**, 3103–3108 (2011).
16. M. Giordano, J. Beardall, J. A. Raven, CO2 concentrating mechanisms in algae: Mechanisms, environmental modulation, and evolution. *Annu. Rev. Plant Biol.* **56**, 99–131 (2005).
17. P. D. Tortell, G. H. Rau, F. M. M. Morel, Inorganic carbon acquisition in coastal Pacific phytoplankton communities. *Limnol. Oceanogr.* **45**, 1485–1500 (2000).
18. E. Derelle, C. Ferraz, S. Rombauts, P. Rouzé, A. Z. Worden, S. Robbens, F. Partensky, S. Degroeve, S. Echeynié, R. Cooke, Y. Saeys, J. Wuyts, K. Jabbari, C. Bowler, O. Panaud, B. Piégu, S. G. Ball, J.-P. Ral, F.-Y. Bouget, G. Piganeau, B. De Baets, A. Picard, M. Delseny, J. Demaille, Y. Van de Peer, H. Moreau, Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 11647–11652 (2006).
19. A. Z. Worden, J.-H. Lee, T. Mock, P. Rouzé, M. P. Simmons, A. L. Aerts, A. E. Allen, M. L. Cuvelier, E. Derelle, M. V. Everett, E. Foulon, J. Grimwood, H. Gundlach, B. Henrissat, C. Napoli, S. M. McDonald, M. S. Parker, S. Rombauts, A. Salamov, P. Von Dassow, J. H. Badger, P. M. Coutinho, E. Demir, I. Dubchak, C. Gentemann, W. Eikrem, J. E. Gready, U. John, W. Lanier, E. A. Lindquist, S. Lucas, K. F. X. Mayer, H. Moreau, F. Not, R. Otillar, O. Panaud, J. Pangilinan, I. Paulsen, B. Piegu, A. Poliakov, S. Robbens, J. Schmutz, E. Toulza, T. Wyss, A. Zelensky, K. Zhou, E. V. Armbrust, D. Bhattacharya, U. W. Goodenough, Y. Van de Peer, I. V. Grigoriev, Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* **324**, 268–272 (2009).
20. Y. Tsuji, I. Suzuki, Y. Shiraiwa, Enzymological evidence for the function of a plastid-located pyruvate carboxylase in the haptophyte alga *Emiliania huxleyi*: A novel pathway for the production of C_4 compounds. *Plant Cell Physiol.* **53**, 1043–1052 (2012).
21. X. L. Shi, D. Marie, L. Jardillier, D. J. Scanlan, D. Vaulot, Groups without cultured representatives dominate eukaryotic picophytoplankton in the oligotrophic South East Pacific Ocean. *PLOS ONE* **4**, 11 (2009).
22. S. Pesant, F. Not, M. Picheral, S. Kandels-Lewis, N. Le Bescot, G. Gorsky, D. Iudicone, E. Karsenti, S. Speich, R. Troublé, C. Dimier, S. Searson, Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data* **2**, 150023 (2015).
23. S. Sunagawa, S. G. Acinas, P. Bork, C. Bowler, D. Eveillard, G. Gorsky, L. Guidi, D. Iudicone, E. Karsenti, F. Lombard, H. Ogata, S. Pesant, M. B. Sullivan, P. Wincker, C. de Vargas, Tara Oceans: Towards global ocean ecosystems biology. *Nat. Rev. Microbiol.* **18**, 428–445 (2020).
24. C. M. Duarte, Seafaring in the 21St century: The Malaspina 2010 circumnavigation expedition. *Limnol. Oceanogr. Bull.* **24**, 11–14 (2015).
25. S. J. Biller, P. M. Berube, K. Dooley, M. Williams, B. M. Satinsky, T. Hackl, S. L. Hogle, A. Coe, K. Bergauer, H. A. Bouman, T. J. Browning, D. De Corte, C. Hassler, D. Hulston, J. E. Jacquot, E. W. Maas, T. Reinthaler, E. Sintes, T. Yokokawa, S. W. Chisholm, Marine microbial metagenomes sampled across space and time. *Sci. Data* **5**, 180176 (2018).
26. L. P. Coelho, R. Alves, Á. R. del Río, P. N. Myers, C. P. Cantalapiedra, J. Giner-Lamia, T. S. Schmidt, D. R. Mende, A. Orakov, I. Letunic, F. Hildebrand, T. Van Rossum, S. K. Forslund, S. Khedkar, O. M. Maistrenko, S. Pan, L. Jia, P. Ferretti, S. Sunagawa, X.-M. Zhao, H. B. Nielsen, J. Huerta-Cepas, P. Bork, Towards the biogeography of prokaryotic genes. *Nature* **601**, 252–256 (2022).
27. A. Minhas, B. Kaur, J. Kaur, Genomics of algae: Its challenges and applications, in *Pan-Genomics: Applications, Challenges, and Future Prospects* (Elsevier, 2020), pp. 261–283.
28. G. W. Tyson, J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, J. F. Banfield, Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
29. T. O. Delmont, M. Gaia, D. D. Hinsinger, P. Frémont, C. Vanni, A. Fernandez-Guerra, A. M. Eren, A. Kourlaiev, L. d'Agata, Q. Clayssen, E. Villar, K. Labadie, C. Cruaud, J. Poulain, C. Da Silva, M. Wessner, B. Noel, J.-M. Aury, S. Sunagawa, S. G. Acinas, P. Bork, E. Karsenti, C. Bowler, C. Sardet, L. Stemmann, C. de Vargas, P. Wincker, M. Lescot, M. Babin, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, O. Jaillon, S. Kandels, D. Iudicone, H. Ogata, S. Pesant, M. B. Sullivan, F. Not, K.-B. Lee, E. Boss, G. Cochrane, M. Follows, N. Poulton, J. Raes, M. Sieracki, S. Speich, C. de Vargas, C. Bowler, E. Karsenti, E. Pelletier, P. Wincker, O. Jaillon, Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics* **2**, 100123 (2022).
30. A. Peterson, J. Soberón, Species distribution modeling and ecological niche modeling: Getting the concepts right. *Nat. Conservação* **10**, 1–6 (2012).
31. F. T. Dahlke, S. Wohlrab, M. Butzin, H.-O. Pörtner, Thermal bottlenecks in the life cycle define climate vulnerability of fish. *Science* **369**, 65–70 (2020).
32. G. Beaugrand, A. Conversi, A. Atkinson, J. Cloern, S. Chiba, S. Fonda-Umani, R. R. Kirby, C. H. Greene, E. Goberville, S. A. Otto, P. C. Reid, L. Stemmann, M. Edwards, Prediction of unprecedented biological shifts in the global ocean. *Nat. Clim. Change* **9**, 237–243 (2019).
33. E. Faure, S.-D. Ayata, L. Bittner, Towards omics-based predictions of planktonic functional composition from environmental data. *Nat. Commun.* **12**, 4361 (2021).
34. P. Frémont, M. Gehlen, M. Vrac, J. Leconte, T. O. Delmont, P. Wincker, D. Iudicone, O. Jaillon, Restructuring of plankton genomic biogeography in the surface ocean under climate change. *Nat. Clim. Change* **12**, 393–401 (2022).
35. D. J. Richter, R. Watteaux, T. Vannier, J. Leconte, P. Frémont, G. Reygondeau, N. Maillet, N. Henry, G. Benoit, O. Da Silva, T. O. Delmont, A. Fernàndez-Guerra, S. Suweis, R. Narci, C. Berney, D. Eveillard, F. Gavory, L. Guidi, K. Labadie, E. Mahieu, J. Poulain, S. Romac, S. Roux, C. Dimier, S. Kandels, M. Picheral, S. Searson, T. O. Coordinators, S. Pesant, J.-M. Aury, J. R. Brum, C. Lemaitre, E. Pelletier, P. Bork, S. Sunagawa, F. Lombard, L. Karp-Boss, C. Bowler, M. B. Sullivan, E. Karsenti, M. Mariadassou, I. Probert, P. Peterlongo, P. Wincker, C. de Vargas, M. Ribera d'Alcalà, D. Iudicone, O. Jaillon, Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems. *eLife* **11**, e78129 (2022).
36. L. Nespoli, V. Medici, Multivariate boosted trees and applications to forecasting and control. *J. Mach. Learn. Technol.* **23**, 47 (2022).
37. N. Grimsley, S. Yau, G. Piganeau, H. Moreau, Typical features of genomes in the Mamiellophyceae, in *Marine Protists: Diversity and Dynamics*, S. Ohtsuka, T. Suzaki, T. Horiguchi, N. Suzuki, F. N, Eds. (Springer, 2015), pp. 107–127.
38. M. Haimovich-Dayan, N. Garfinkel, D. Ewe, Y. Marcus, A. Gruber, H. Wagner, P. G. Kroth, A. Kaplan, The role of C_4 metabolism in the marine diatom *Phaeodactylum tricornutum*. *New Phytol.* **197**, 177–185 (2013).
39. P. J. McGinn, F. M. M. Morel, Expression and inhibition of the carboxylating and decarboxylating enzymes in the photosynthetic C_4 pathway of marine diatoms. *Plant Physiol.* **146**, 300–309 (2008).
40. G. Piganeau, N. Grimsley, H. Moreau, Genome diversity in the smallest marine photosynthetic eukaryotes. *Res. Microbiol.* **162**, 570–577 (2011).
41. A. S. Rigual-Hernández, T. W. Trull, J. A. Flores, S. D. Nodder, R. Eriksen, D. M. Davies, G. M. Hallegraeff, F. J. Sierro, S. M. Patil, A. Cortina, A. M. Ballegeer, L. C. Northcote, F. Abrantes, M. M. Rufino, Full annual monitoring of Subantarctic *Emiliania huxleyi* populations reveals highly calcified morphotypes in high-CO_2 winter conditions. *Sci. Rep.* **10**, 2594 (2020).
42. R. Clement, E. Jensen, L. Prioretti, S. C. Maberly, B. Gontero, Diversity of CO_2-concentrating mechanisms and responses to CO_2 concentration in marine and freshwater diatoms. *J. Exp. Bot.* **68**, 3925–3935 (2017).
43. M. J. Behrenfeld, K. H. Halsey, A. J. Milligan, Evolved physiological responses of phytoplankton to their integrated growth environment. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 2687–2703 (2008).
44. B. A. Ward, S. Dutkiewicz, O. Jahn, M. J. Follows, A size-structured food-web model for the global ocean. *Limnol. Oceanogr.* **57**, 1877–1891 (2012).
45. X. Yin, P. C. Struik, Exploiting differences in the energy budget among C_4 subtypes to improve crop productivity. *New Phytol.* **229**, 2400–2409 (2021).
46. M. J. Behrenfeld, O. Prasil, Z. S. Kolber, M. Babin, P. G. Falkowski, Compensatory changes in Photosystem II electron turnover rates protect photosynthesis from photoinhibition. *Photosynth. Res.* **58**, 259–268 (1998).
47. A. Duncan, K. Barry, C. Daum, E. Eloe-Fadrosh, S. Roux, K. Schmidt, S. G. Tringe, K. U. Valentin, N. Varghese, A. Salamov, I. V. Grigoriev, R. M. Leggett, V. Moulton, T. Mock, Metagenome-assembled genomes of phytoplankton microbiomes from the Arctic and Atlantic Oceans. *Microbiome* **10**, 67 (2022).

48. J. Leconte, L. F. Benites, T. Vannier, P. Wincker, G. Piganeau, O. Jaillon, Genome resolved biogeography of Mamiellales. *Genes* **11**, 66 (2020).
49. C. Lovejoy, W. F. Vincent, S. Bonilla, S. Roy, M.-J. Martineau, R. Terrado, M. Potvin, R. Massana, C. Pedrós-Alió, Distribution, phylogeny, and growth of cold-adapted Picoprasinophytes in Arctic Seas1. *J. Phycol.* **43**, 78–89 (2007).
50. N. Trefault, R. De la Iglesia, M. Moreno-Pino, A. Lopes dos Santos, C. Gérikas Ribeiro, G. Parada-Pozo, A. Cristi, D. Marie, D. Vaulot, Annual phytoplankton dynamics in coastal waters from Fildes Bay, Western Antarctic Peninsula. *Sci. Rep.* **11**, 1368 (2021).
51. M. Marquardt, A. Vader, E. I. Stübner, M. Reigstad, T. M. Gabrielsen, Strong seasonality of marine microbial eukaryotes in a high-Arctic Fjord (Isfjorden, in West Spitsbergen, Norway). *Appl. Environ. Microbiol.* **82**, 1868–1880 (2016).
52. A. R. Kirkham, C. Lepère, L. E. Jardillier, F. Not, H. Bouman, A. Mead, D. J. Scanlan, A global perspective on marine photosynthetic picoeukaryote community structure. *ISME J.* **7**, 922–936 (2013).
53. A. R. Kirkham, L. E. Jardillier, R. Holland, M. V. Zubkov, D. J. Scanlan, Analysis of photosynthetic picoeukaryote community structure along an extended Ellett Line transect in the northern North Atlantic reveals a dominance of novel prymnesiophyte and prasinophyte phylotypes. *Deep-Sea Res. I Oceanogr. Res. Pap.* **58**, 733–744 (2011).
54. A. Meng, E. Corre, I. Probert, A. Gutierrez-Rodriguez, R. Siano, A. Annamale, A. Alberti, C. Da Silva, P. Wincker, S. Le Crom, F. Not, L. Bittner, Analysis of the genomic basis of functional diversity in dinoflagellates using a transcriptome-based sequence similarity network. *Mol. Ecol.* **27**, 2365–2380 (2018).
55. H. J. Atkinson, J. H. Morris, T. E. Ferrin, P. C. Babbitt, Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLOS ONE* **4**, e4345 (2009).
56. T. P. Boyer, H. E. Garcia, R. A. Locarnini, M. M. Zweng, A. V. Mishonov, J. R. Reagan, K. A. Weathers, O. K. Baranova, D. Seidov, I. V. Smolyar, *World Ocean Atlas 2018* (National Centers for Environmental Information, 2018); www.ncei.noaa.gov/archive/accession/NCEI-WOA18.
57. A. Morel, S. Maritorena, Bio-optical properties of oceanic waters: A reappraisal. *J. Geophys. Res. Oceans* **106**, 7163–7180 (2001).
58. C. de Boyer Montégut, G. Madec, A. S. Fischer, A. Lazar, D. Iudicone, Mixed layer depth over the global ocean: An examination of profile data and a profile-based climatology. *J. Geophys. Res. Oceans* **109**, doi.org/10.1029/2004JC002378 (2004).
59. T. Aramaki, R. Blanc-Mathieu, H. Endo, K. Ohkubo, M. Kanehisa, S. Goto, H. Ogata, KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
60. M. Manni, M. R. Berkeley, M. Seppey, E. M. Zdobnov, BUSCO: Assessing genomic data quality and beyond. *Curr. Protoc.* **1**, e323 (2021).
61. Y. Escoufier, *Echantillonnage dans une Population de Variables Aleatoires Reelles*. (Dept. de math.; Univ. des sciences et techniques du Languedoc, Montpellier, 1970).
62. D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, C. F. Dormann, Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**, 913–929 (2017).
63. T. O. Delmont, M. Gaia, D. D. Hinsinger, P. Fremont, C. Vanni, A. F. Guerra, A. M. Eren, A. Kourlaiev, L. d'Agata, Q. Clayssen, E. Villar, K. Labadie, C. Cruaud, J. Poulain, C. Da Silva, M. Wessner, B. Noel, J.-M. Aury; Tara Oceans Coordinators, C. de Vargas, C. Bowler, E. Karsenti, E. Pelletier, P. Wincker, O. Jaillon, Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics. [Preprint] (2020). https://doi.org/10.1101/2020.10.15.341214.
64. K. Rizzolo, S. E. Cohen, A. C. Weitz, M. M. López Muñoz, M. P. Hendrich, C. L. Drennan, S. J. Elliott, A widely distributed diheme enzyme from Burkholderia that displays an atypically stable bis-Fe(IV) state. *Nat. Commun.* **10**, 1101 (2019).
65. D. Forster, L. Bittner, S. Karkar, M. Dunthorn, S. Romac, S. Audic, P. Lopez, T. Stoeck, E. Bapteste, Testing ecological theories with sequence similarity networks: Marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. *BMC Biol.* **13**, 16 (2015).
66. F. Murtagh, P. Legendre, Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *J. Classif.* **31**, 274–295 (2014).

Acknowledgments: The coauthors wish to thank public taxpayers who fund their salaries. Salaries of A.S. and P.D. were financed by the Blue-Cloud European project (grant agreement no. 862409). The authors want to thank all the people involved in the Tara Oceans project for making data publicly available. **Funding:** This work was supported by the European Union's Horizon program (call BG-07-2019-2020, topic: [A] 2019 - Blue-Cloud services, grant agreement no. 862409). **Author contributions:** Writing—original draft: A.S., L.G., and J.O.I. Conceptualization: A.S., P.D., S.-D.A., L.B., E.P., L.G., and J.O.I. Investigation: A.S., P.D., L.B., E.P., and L.G. Writing—review and editing: A.S., S.-D.A., E.P., L.G., and J.O.I. Methodology: A.S., P.D., L.B., E.P., and J.O.I. Funding acquisition: L.G. and J.O.I. Resources: P.D., L.B., E.P., L.G., and J.O.I. Data curation: A.S., P.D., L.B., E.P., and J.O.I. Validation: A.S., P.D., L.B., E.P., and J.O.I. Supervision: L.G. and J.O.I. Formal analysis: A.S., P.D., L.B., E.P., and J.O.I. Software: A.S., P.D., L.B., E.P., and J.O.I. Project administration: L.G. and J.O.I. Visualization: A.S., S.-D.A., E.P., and J.O.I. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 29 September 2023
Accepted 11 July 2024
Published 16 August 2024
10.1126/sciadv.adl0534