Microbial nitrogen guilds in the oceans: from the bathypelagic to the ocean surface

Juan Rivas-Santisteban[1], Nuria Fernández-González[1,2†], Rafael Laso-Pérez[1,3†], Javier Tamames[1*], Carlos Pedrós-Alió[1*]

[1]Systems Biology Department, Centro Nacional de Biotecnología (CNB-CSIC), C. Darwin nº 3, Madrid, 28049, Spain.
[2]Facultad de Ciencias Experimentales, Universidad Francisco de Vitoria, Ctra. Pozuelo-Majadahonda Km 1,800, Pozuelo de Alarcón, 28223, Spain.
[3]Biogeochemistry and Microbial Ecology Department, Museo Nacional de Ciencias Naturales (MNCN-CSIC), C. de José Gutiérrez Abascal, Madrid, 28006, Spain.

*Corresponding author(s). E-mail(s): jtamames@cnb.csic.es; cpedros@cnb.csic.es;
Contributing authors: jrisant@outlook.es; nuria.fernandez@ufv.es; rafael.laso@mncn.csic.es;
†These authors contributed equally to this work.

Abstract

Microbial nitrogen guilds play crucial roles in the oceanic nitrogen cycle, but their composition and importance across specific depths and conditions remain unclear. In this study, we examine the guilds participating in nitrogen transformations within the low and mid-latitude ocean ecosystems, from the surface down to 4000 m, using data obtained from 75 samples belonging to 11 stations in the Malaspina (MP) dataset. In particular, we used a quantitative approach to investigate the stability of nitrogen acquisition and nitrogen-redox guilds separately. Our results showed that nitrogen acquisition guilds are more stable and redundant than nitrogen-redox guilds across depths and conditions related to specific water columns. However, differential conditions such as nitrogen depletion and oxygen availability affected the two groups of guilds in different ways. These findings have significant implications for the understanding of global nutrient fluxes and the biosphere's functional diversification.

Keywords: microbial guilds, marine ecology, nitrogen cycle, microbial ecology, protein diversification

1 Introduction

Microorganisms have colonized a multitude of niches distributed over the Earth's superficial layers, including the lithosphere, the hydrosphere, and the lower layers of the atmosphere [1]. The functional repertoire found in them has diversified to non obvious extents, making the study of nutrient fluxes very hard [2–4]. Among them, water bodies represent the most important biological functional reservoirs in terms of global nutrient fluxes [5–7]. In consequence, we may still be unaware of fundamental biotic processes in the oceans. For example, anaerobic ammonia oxidation mediated by microbes (anammox) is an ecologically very relevant process ($\approx \frac{1}{2}$ of total N_2 release from the oceans is attributed to anammox [8], while oscillating between 4-79% in specific conditions [9]), but, surprisingly, it had not been uncovered until three decades ago [10].

Part of the functional variation in the oceans is explained by the fact that the Earth's water bodies do not constitute a single biome, but rather a compendium of them corresponding to variable characteristics such as salinity, temperature, access to nutrients (including irradiance), or hydrostatic pressure [11]. To functionally study microbiomes across these ever-changing environments, we propose using a suitable quantitative approach such as microbial guilds [12]. In the present contribution, microbial guilds belonging to low and mid-latitude oceanic biomes were analyzed, as represented by metagenomes from the Malaspina 2010 (MP) expedition [13, 14]. For this purpose, we classified the MP metagenomes according to their corresponding depth, establishing three categories: epipelagic (0-200 m depth), mesopelagic (200-1000 m depth) and bathypelagic (1000-4000 m depth). This depth criterion is typically applied to understand marine ecology because these discretized oceanic layers share many physicochemical characteristics [15, 16]. However, geographic location is also a key factor to consider when establishing the oceanic physicochemical regions, as the Longhurst's provinces attempt to do [17]. Therefore, we considered both depth and specific locations to study how they influence the oceanic microbial functions.

In particular, we focus on the main guilds of the N cycle. The reasons for this are manifold: (i) there exist up to eight redox states in the range of possible N compounds in the ocean [18]; (ii) they consist of both organic and inorganic N species [19]; (iii) N compounds are used in a variety of ways, both as an energy source and as an essential nutrient, and can sometimes serve as an electron acceptor under suboxic conditions [20]; (iv) there is also variability in N-limitation among different areas and compounds, for example, N_2 is never limiting, whereas nitrate (NO_3-) frequently limits phytoplankton growth [18, 20]; and (v) the main functions of the N cycle and its substrates are well described [21–23]. Thus, the N cycle lies in an intermediate stage between complexity and predictability, making this an interesting candidate to describe the change in its functions in detail.

Just as we have classified different MP environments using the depth-location criterion, we'll also classify the guilds related to the N cycle according to the kind of ecological processes in which they are involved. We established the two most general categories: (1) acquisition of N compounds, and (2) redox transformation of the N species acquired from the environment.

Regarding (1), we already pointed out that there are a lot of different N compounds that microorganisms can import from oceanic waters. Within (2) we find transformation processes where imported N compounds act as electron donors or acceptors. To clarify the dynamics of N uptake and N redox transformations, and determining how they respond to depth, is something that has been tentatively explored [14, 24, 25], but it is not yet described using a quantitative ecological framework. With this clear division between biomes and guilds, their interplay in the MP metagenomes will be described.

2 Results

2.1 N/P ratios of the main MP water columns

In order to look for differences between the contributions of different guilds to the N cycle, we already considered the different depth categories as environmental factors, but what geographical criteria could we use that significantly affected N-related functions? We thought of the N/P ratio, which is an environmental factor of interest for oceanographic characterization. It provides a rough idea of how limiting is the dissolved N in an environmental context [26]. When we calculated this ratio in our samples, from the mesopelagic to the deepest, the values stabilized. We observed that the values were only slightly different from 16:1 with depth – the Redfield ratio which is considered not to be either N or P limiting (Fig. 1). However, two stations showed markedly different profiles. The first was a water column off the coast of Mexico (corresponding to station 120), where an Oxygen Minimum Zone (OMZ) occurs. This station had a very low mesopelagic maximum (N/P = 13.86), and its values were always lower than the samples from other stations below a certain depth, as shown in Figure 1. The second was a station from the open North Atlantic Ocean (corresponding to the station 141), whose mesopelagic maximum was 18.96 and its values were always higher than the other stations below this point.

In summary, station 120 reached the mesopelagic N/P maximum faster with depth, close to $\approx$ 210m, while all the other stations showed the maximum at $>$ 400m. Station 141, on the contrary, reached the mesopelagic maximum at the mean depth, but after that point it was never a nitrogen-scarce environment (N/P $\approx$ 16). Therefore, we selected these two water columns to study further changes in N guilds, as they showed opposite departures from the general pattern found regarding N depletion.

2.2 Taxonomic and depth distribution of N functional markers

We determined the abundance of selected N functional markers in the 75 samples (Fig. 2) as curated normalized metagenomic counts (see Methods). The taxonomic composition is shown in Fig. 2a–c in a relative way so that the composition of the less abundant markers can be appreciated. The sum of all the counts from all the samples is shown in Fig 2d–e to compare their net abundance. The most abundant genes were *amt* (marker for ammonia transport), *ilvC* (marker for branched aminoacid synthesis), *nirK* (marker for denitrification and hydroxylamine oxidation in archaea [27]), and *ureC* (marker for urea hydrolysis) regardless of depth. There was considerable

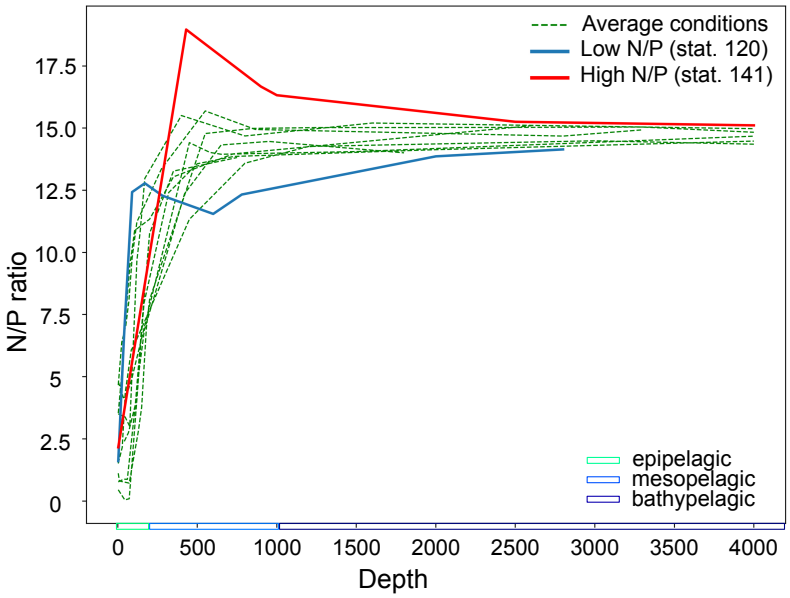

Fig. 1 Different trends in N/P ratios observed in the main MP water columns (11 stations). The bathypelagic depth makes the N/P ratio similar between markedly different stations. Of these, station 120 (Pacific, off the Mexico coast, with a reported oxygen minimum) and station 141 (North Atlantic, open ocean) have the most inconsistent profiles.

species turnover between the epipelagic and mesopelagic, but the latter and the bathypelagic were similar. There was predominance of primary producers (*Cyanobacteria*) at the surface, but it was replaced by *Nitrososphaeria* at depth. As a consequence, the *nirK* gene became more abundant than *amt*. The marker for N-fixation (*nifH*) was only found a few times with extremely few counts and, therefore, it does not appear in Figure 2d–f. However, we showed the relative proportion of the few counts in Figure 2a–c. Some functional markers were found only at some depths. For example, ammonia oxidation (*amoA*) was absent from the epipelagic but it was substantial at the other depths. Likewise, there was a maximum of *nosZ* (marker for the last step of denitrification) in the bathypelagic and *Actinobacteria* appeared at these depths, contributing moderately to several functions.

In summary (1), as expected, the most popular implementation for inorganic N uptake function in the oceans was the ammonium transporter, (2) the preferred metabolic usage of organic-N was to build branched chain amino acids regardless of depth, and (3) the main redox process involving N species in the oceans concerned *nirK* functions (denitrification / nitrification partial steps), particularly below the euphotic zone. In the next section we describe spatial and compositional trends of the main guilds.

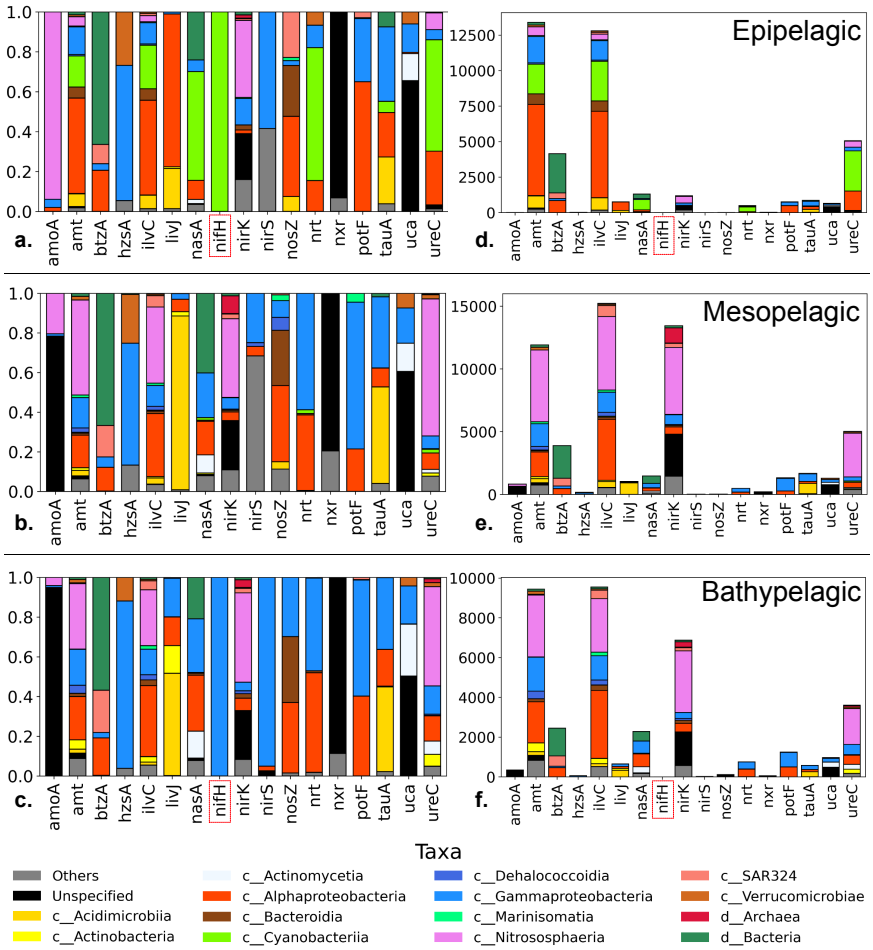

Fig. 2 Relative proportion (a–c) and sum (d–f) of normalized counts (TPM) for the main N functional markers in the tropical and subtropical ocean. All the graphs shown have removed sequences not considered to carry out the assigned function. After filtering, N fixation appears only very lightly in the epipelagic and bathypelagic (indicated in red), implemented in distinct taxa. The quantity and type of these sequences were detailed in Figure S1. The most detectable changes occurred between the epipelagic (top) and mesopelagic (middle), and were minor from the latter to the bathypelagic (bottom). These changes included a threefold net increase for nitrite reduction (*nirK*) and ammonium oxidation (*amoA*) markers abundance, as well as a significant decrease in urease (*ureC*) and amino acid synthesis (*ilvC*) markers. Furthermore, these changes were also reflected in the taxonomic composition (presence of *Cyanobacteria* in the euphotic zone and the dominance of *Nitrososphaeria* from the mesopelagic to the bathypelagic).

2.3 Description of N uptake guilds

Based on the information in figures 1 and 2, we studied separately four differential environments: OMZ with anammox off the Mexican West coast (corresponding with the blue line in Fig. 1), the open Atlantic Ocean water column (corresponding with the red line in Fig. 1), the non-OMZ near Brazil East coast water column (mean

N/P conditions, but associated with an abnormal presence of *Alcanivorax* spp.), and the average MP conditions. Then, we examined all three depths levels for each case: epipelagic, mesopelagic and bathypelagic; making a total of 12 contexts (4 x 3). We estimated the importance of functions as a product between normalized abundances and non-expected richness for the relevant N uptake functional markers (Fig. 3; see Methods). This metric is known as k-value [12]. The k-values were log transformed and displayed in radial plots for clarity when identifying patterns between contexts.

First, we found that patterns of individual taxonomic contributors to the general N uptake guild changed both among sampling stations and among depths. However, environments belonging to the same ocean layer often showed taxonomical identity. Thus, we realized that N acquisition was dominated by specific taxa in the epipelagic, regardless of the geographical location of the water column. These were mainly *Pseudomonas taenensis*, *Prochlorococcus marinus*, *Sulfitobacter pontiacus*, *Alteromonas macleodii*, and *Mesorhizobium oceanicum*. Regarding NO_3- acquisition (inferred by the presence of the substrate-binding subunit of the nitrate transporter, *nrtA*) the most important species were *Prochlorococcus marinus* and *Sulfitobacter pontiacus*, and *Alteromonas macleodii*, with some *Alcanivorax* spp. contributing in the SAtl station. Regarding organic N acquisition, the outstanding contributors were members of *Pseudomonadales* and *Acidimicrobiia*, while some belonged to the *Desulfobacteraceae* family (UBA 2156).

With depth, *Nitrososphaeria* dominated the inorganic N uptake via ammonia import, as shown in Figure 2. However, regarding the import other N compounds, this group was not found to be relevant. In the mesopelagic, the general N uptake guild pattern (bottom row of Fig. 3) was remarkably similar to the epipelagic, slightly increasing importance over non-AA uptake, but decreasing importance considering the rest. However, in given water columns, we found several differences, which can be related to specific conditions. Most differences occurred regarding NO_3- import, where there was absence in mesopelagic OMZ, while in the higher N/P water column it was inferior but similar to mean conditions. Finally, in the SAtl station, we found a remarkable increase in *nrtA* relevance, mainly due to the *Alcanivorax* sp002354605 and sp000155615 abnormal presence. There are other differences in the acquisition of taurine, polyamines and AAs, but these are of lesser importance. In general, *Sulfitobacter pontiacus* was found to be a key contributor in terms of nitrate and putrescine acquisition from the mesopelagic downwards, although neither in the OMZ water column nor in the NAtl mesopelagic.

As for the bathypelagic, we found that patterns were more similar between stations, but it was the most different pattern in terms of depth. However, top N importers were less diverse. NO_3- acquisition increased in all cases except for SAtl station, while general AA acquisition was partially replaced by branched AA acquisition. It was also the only depth at which the import function of sulphonated compounds was relevant (*ssuA*; taurine is a shared substrate between Ssu and Tau ABC transporters), which can serve as a source of N and S. However, regarding Non AA organic N assimilation, putrescine uptake was of higher importance at this layer.

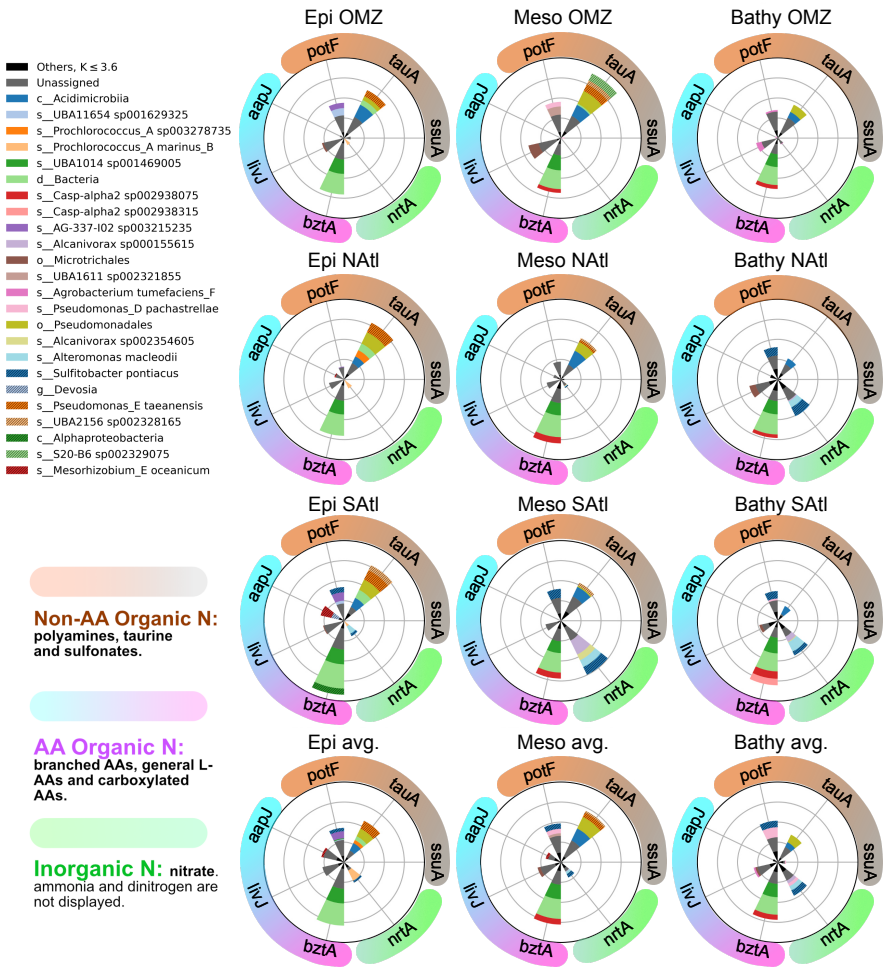

Fig. 3 N uptake guild patterns in the tropical and subtropical oceans represented by log(k-value) [12]. We represented the top 25 taxonomic contributors, for the sake of clarity. All other taxa were grouped under "Others". The general pattern of N acquisition genes was similar in all the stations (last row, MP average). Ammonia uptake was always important, as *amt* was always the most abundant gene, showing the preference of most bacterioplankton for ammonia. Therefore, we did not represented it. Next in importance were the organic nitrogen transporters for carboxylated amino-acids (*bztA*) and for taurine (*tauA*). The remaining organic-N transporters were of minor importance. The marker gene for nitrate uptake (*nrtA*) was absent from the OMZ station and was present only at the bathypelagic in the NAtl station. However, it was very important at the SAtl station, particularly at the mesopelagic.

2.4 Description of N-redox transformation guilds

Using the same environmental classification, we studied N-redox functional relationships between species (Fig. 4). In this case changes were very marked between depths and between stations. In general, the two Atlantic stations behaved similarly to the ocean average (bottom of Fig. 4), but with lower species and function richness.

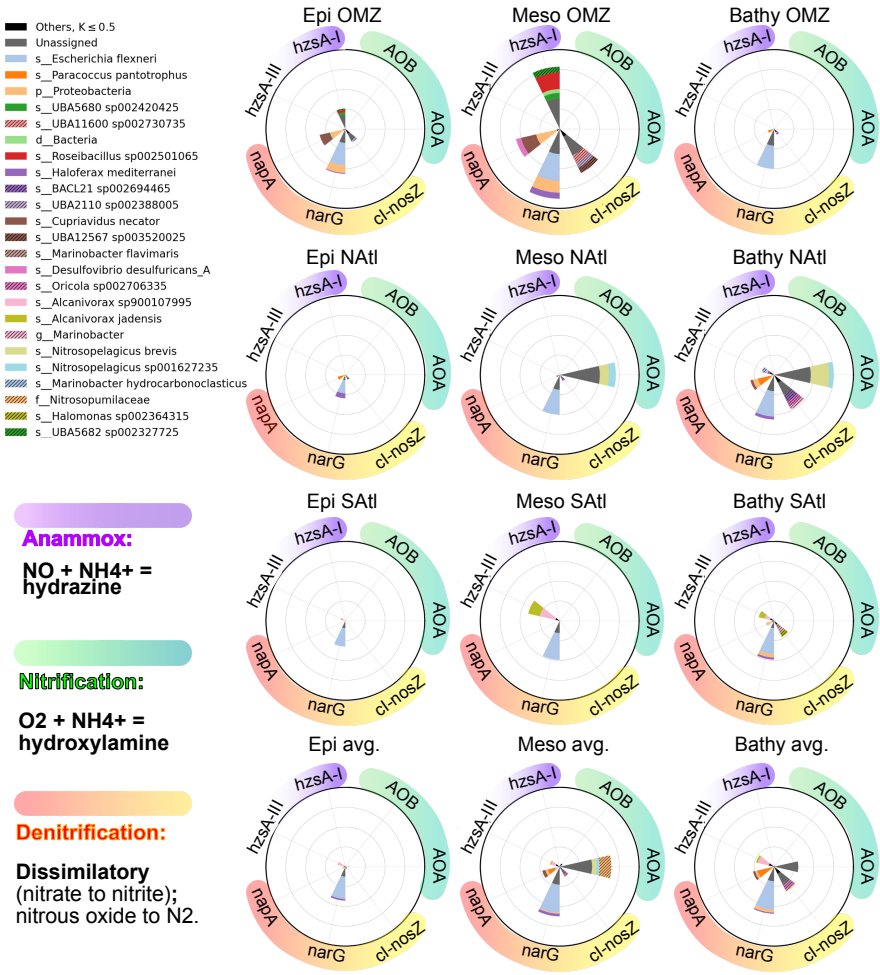

Fig. 4 The N redox guild patterns in the tropical and subtropical oceans. Unlike with the nitrogen acquisition genes, those involved in redox processes varied substantially from one station to another. Anammox (reflected by *hzsA*-I), as expected, was important only in the OMZ station in the mesopelagic. Denitrification followed the same pattern as in the global ocean (last row) with *narG* as the most important gene and increasing with depth, while being similar between the mesopelagic and the bathypelagic. The exception was the OMZ station, where the three examined genes were more important at the surface and the mesopelagic. Nitrification was surprisingly absent from both the OMZ and the SAtl station.

In general, ammonia oxidation was relevant in the mesopelagic, mainly performed by *Nitrosopelagicus brevis* and related archaea. Denitrification first step (*narG*) was relevant in all stations. However, both denitrification and nitrification became more important with depth, and were comparatively weak in the epipelagic. This fitted well with our environmental data, where the epipelagic always represented a N scarcity scenario. However, station 120 (OMZ) showed a different pattern. Anammox (*hzsA-I*; a marker for hydrazine synthesis, main step in anammox, which requires the products

Table 1 Generalized Linear Model of *delta* values for protein diversification.

Context	$\text{coef}_{N_{uptake}}$	$\text{std. err}_{N_{uptake}}$	$P > \lvert z \rvert_{N_{uptake}}$	$\text{coef}_{N_{redox}}$	$\text{std. err}_{N_{redox}}$	$P > \lvert z \rvert_{N_{redox}}$
OMZ	-0.239	0.078	0.002 ***	-0.203	0.075	0.007 ***
NAtl	-0.105	0.078	0.178	-0.0891	0.091	0.326
SAtl	-0.1555	0.082	0.059	-0.135	0.108	0.212
Epi	0.17	0.103	0.098	0.222	0.119	0.062
Meso	0.262	0.102	0.01 **	0.33	0.117	0.005 ***
Bathy	0.432	0.103	0.000 ****	0.268	0.118	0.023 *

of *nirS* activity – NO and ammonia –, and very low O_2 concentrations) was only present under OMZ conditions. Here, denitrification markers were also most abundant, with strong presence of *nosZ*, the last step.

In the SAtl water column, we found a very surprising change: there was no ammonium oxidation in the mesopelagic, but instead we reported a type of *hzsA* that will only be relevant in aerobic environments (*hzsA*-III). We suggest that this type of protein sequences corresponds to a neofunctionalization [28] of the *hzsA* gene (Fig. S2), since (1) anammox function is only possible under low O_2 tension (upper limit explored to be 20 μmol/L [29]) and here it was found at a concentration of 191.52 μmol/L; (2) hydrazine synthase activity is thought to be bound to the anammoxosome, an organelle not present in *Alcanivorax* [30]; and (3) this type of protein sequence has never been found where ideal conditions for anammox occur in our dataset (0.89 μmol/L O_2, in samples from the OMZ water column); instead, the orthodox *hzsA*-I associated with *Planctomycetota* was found with a remarkable relevance.

We also concluded that, in contrast to N uptake guilds, N-redox guilds seemed to be more taxonomically linked, as we never identified different markers implemented in the same taxa among the top contributors.

2.5 Diversification of N cycle proteins

Functions can be carried out by a single protein sequence or by a multitude of equivalent sequence variants. To study how protein diversification increased or decreased depending on the environment, we calculated a ratio between observed and expected unique sequence richness, *delta* (Methods). Then, we compared *delta* values of the previously defined functional categories: proteins that serve for the acquisition of different N compounds and proteins that transform these compounds by redox processes.

Using a generalized linear model, we found that the location of individual samples did not explain significant *delta* variance in either protein type, except for the OMZ (Table 1). We found that both N uptake and N redox proteins underdiversified at this station, althought their abundance did not differed from the average (Figs. S3, S4 & S5). This correspondence between *delta* variation and individual stations was only found in this water column, regardless of the protein function type. In contrast, we found a highly significant tendency for N redox-related proteins to diversify above the expected in the ocean mesopelagic and bathypelagic average (Table 1).

3 Discussion

3.1 Are functional compartiments in the oceans taxonomically linked?

Recent research has improved our understanding of marine microbial functions and their role in biogeochemical cycling, as well as helping us to identify the main players driving these processes [31–33]. First of all, our results yielded novel data on the taxonomic composition of the oceanic N guilds, which remained, to some extent, compartmentalized throughout the water column (Fig. 2). For example, *Nitrososphaeria* relevant functions remained very abundant with depth in the acquisition of ammonia – which most of it will be ultimately oxidized –, but not in the epipelagic (Fig. 2). However, vertical stratifiction of different taxa is not a new concept in marine microbial ecology [34]. For example, we already knew of specific *Prochlorococcus* populations linked to the bottom of the photic zone, or some *Crenarachaeota* to be important regarding ammonia oxidation in sub-mesopelagic depths [35]. Likewise, the dynamics of particles – the so-called *marine snow* – are of great importance (whether they sink, whether they are dissolved, or whether they are large and taxonomically diverse [36, 37]). In addition, the internal characteristics of these particles are also critical to support metabolisms and species across depths [38, 39]. Therefore, it was expected that each taxon will occupy an optimum depth to perform its function.

Surprisingly, we found that top contributors to the N guilds retained more identity with depth than with specific conditions (Figs. 3 & 4). For example, *Desulfovibrio desulfuricans* (strong pink in Fig. 4) was only contributing to the denitrification guild in the station 120 mesopelagic, and it was not found in the others. *Cupriavidus necator* (brown in Fig. 4) represents yet another kind of example, since its function was not bound to the station 120 epipelagic and mesopelagic, but neither to a certain depth – it is also found in some other stations' bathypelagic. We interpreted that the species contribution to the N guilds along the water column was more influenced by resource availability and location contingencies than by depth alone. Thus, we considered that functional compartiments were taxonomically dilluted, albeit with certain depth dependence.

3.2 How N scarcity affected the oceanic N guilds

Depth and N availability are closely related in the ocean. However, this relationship is variable with geographic location, as seen in Figure 1. This is why we paid special attention to determine N/P unusual ratios. Once we identified the two water columns with the most divergent distributions, we analyzed both of them separately. The results were already shown and discussed above, but further conclusions can be inferred regarding N shortage or oversupply. In summary, what we observed was that most N-redox guilds were highly reliant on depth, especially compared to the N uptake guilds. In addition, certain specific features of the water column also altered the N-redox guilds more than the N uptake guilds. Now we will focus on discussing the causes of these differential perturbations regarding only N depletion.

The lower values of N/P corresponded to the OMZ water column. Typically, the scarcity of a resource constrains the type and amount of biological processes that depend on it [40, 41]. Interestingly enough, our results are telling us the opposite, at least within guilds for redox transformation of the available N species (Fig. 4): the importance and variety of processes considerably increased when N (the limiting resource) was scarcer compared to all the other water columns in the mesopelagic. However, one straightforward explanation for the fact that N transformation processes multiplied as N depletion increased is precisely that, in the absence of oxygen as the ultimate electron acceptor, N would be used profusely (*via* denitrification and anammox) to obtain reducing power. Surprisingly, the richness of unique protein sequences for the different N functions in the OMZ was lower than expected for their abundance when compared to the rest of the oceanic environments (Table 1, Fig. S3). This latter result is not explained by a poor taxonomic diversity, as OMZs are considered fairly diverse environments [42].

We interpreted both results together. We conjecture that the diversification of proteins exploiting a resource is a process somehow leveraged by the scarcity and by the need of it. This effect would emerge, of course, from the implied eco-evolutionary mechanisms (whether known or unknown [43–45]).

Finally, it can be stated that the lowest N/P mesopelagic minimum in our dataset (13.86) was not limiting for N transforming guilds, in fact, was a potentiator of them. In addition, N uptake guilds were limited by the fact that the availability of the different N species was biased (highly oxidated N states like NO_3- cannot exist in highly reductive environments). Therefore, nitrate acquisition was not relevant in the OMZ, while reduced forms of N were preferred (taurine, carboxylated amino acids, and ammonia).

On the other hand, we explored the higher N/P ratios in our dataset, corresponding with a NAtl water column (Fig. 1). We reported a slight decrease in denitrification in the mesopelagic compared to the mean oceanic conditions, but an increase in the bathypelagic. The most remarkable thing is that the first step in nitrification is still very important, even more than in the mesopelagic. Ammonia oxidation is usually inhibited in euphotic zones [46] and remains very prevalent in the mesopelagic [47]. In this specific case and with our approach, denitrification and nitrification guilds were at their maximum in the bathypelagic, with a strong correspondence with the higher N/P profile of these samples from station 141. With respect to N acquisition guilds, and unlike the MP average, putrescine and nitrate acquisition were not relevant until bathypelagic, correlating negatively with taurine acquisition.

3.3 Stability of N guilds along the water column

With the introduction of a quantitative approach for guilds, we now have a better chance to investigate whether microbial resource-related functions are stable or not in changing environments. However, despite this improvement, the windows we have for observation are still not sufficient to fully and/or accurately determine the stability of guilds in all aspects. Our ability to determine the stability of N guilds depended on three factors: (1) on the reliability of our methods for functional classification of retrieved metagenomic sequences, (2) on how we measure the ecological contribution of

their members, and (3) on the proper discrimination of the guilds. These factors have been discussed at length in Rivas-Santisteban *et al.* [12], and we therefore followed the methodological guidelines established in this previous study.

Considering the diversification of the specific protein effectors that perform a function to measure the stability of that function seems appropriate. This is motivated by the argument that the greater the number of functionally redundant unique sequences for a given environment, the more sequences will still be able to act within their physiological response when these environmental conditions change. Therefore, effector diversification is congruent with functional stability [48].

Bearing in mind the above, our results indicate that the N acquisition guilds were far more stable and redundant than the N-redox guilds across depths and conditions related to specific water columns (Figs. 3 & 4). Elaborating further, we considered the probability distributions of the diversification coefficients, *delta*, across the twelve contexts analyzed (Fig. S3). It seemed that the specific differential conditions – N depletion, O_2 availability, and depth – affected the two group of guilds distinctively. For example, in the OMZ mesopelagic, N uptake proteins diversified very little, while N redox proteins were closer to the expected *delta* mean. However, if only the average oceanic conditions were considered (all MP samples, bottom row of Fig. S3), these were more similar, although they followed a narrower trend with less variability. When comparing these with the probability distributions considering only the normalized abundance values, we did not find a strict correspondence, especially in the case of N uptake guilds (Fig. S4).

A reasonable interpretation for these results is that the processes of protein *overdiversification* and *underdiversification* are, despite their obscure foundations, modulators of the stability of molecular functions, which operate in response to changing environments, as we have seen in Table 1. Both N redox and N uptake guilds exhibited distinct responses to varying conditions in terms of diversification, as *delta* values showed a variable deviation from the expected (Fig. S5). Thus, functional stability appears to hinge on numerous factors, with each guild potentially exhibiting a unique response based on the extent of protein diversification they have undergone. We suspect that diversification processes are not contingent to the species performance: they are linked to the particular relationship between environment, function, and protein evolution.

4 Conclusions

We have been able to describe novel functional dynamics and key players in the N cycle as a function of depth and some specific water columns with respect to their unique characteristics for the Malaspina dataset. In addition to the above, the guild quantification has allowed us to improve automatic annotation and to outline the following important effects:

1) In the mesopelagic and bathypelagic, *Nitrososphaeria* dominated ammonia oxidation (to nitrite), while overriding some of the functions that were carried out by *Cyanobacteria* in the epipelagic (urea hydrolysis and ammonia uptake).

2) N redox guilds showed greater functional variability with depth and with geographic location than N uptake guilds.

3) The importance of denitrification/nitrification in the bathypelagic was strongly dependent on the availability of N.

4) As discussed, some types of anammox marker proteins (related to hydrazine synthesis) might not have a relevant role in OMZs. We called them *hzsA*-III (corresponding to oxygenic conditions with an increased presence of *Alcanivorax* spp. in the SAtl station), in contrast to *Planctomyces*-like *hzsA*-I proteins (corresponding to anoxygenic conditions).

5) The *over/underdiversification* of proteins responded to very specific environmental variables, those that relax or exacerbate the constraining selection over the effectors of the function that the proteins compromise. A clarifying example is that the proteins with uptake activity are significantly less diverse than expected for each taxon in samples from the OMZ. However, it differed for N redox functions. This result also suggests that the type of function conditioned the richness of the proteins that performed it over the water column.

5 Methods

5.1 Sampling and environmental data collection

During the Malaspina 2010 expedition, conducted from 2010 to 2011, genetic and environmental data were gathered from 37 sampling stations spanning various oceanic regions, including the tropical and subtropical zones of the Atlantic, Indian, and Pacific Oceans, at depths ranging from 3 to 4000 meters. The majority of samples originated from the Atlantic, Pacific, and notably the Indian Ocean, where many parameters were documented for the first time during the expedition. Sampling methods and protocols are outlined in [13]. Given the tropical and subtropical nature of the samples, seasonal climate variations are expected to have minimal impact. Metadata encompassed date, latitude, longitude, depth, temperature, filter size, conductivity, oxygen levels, fluorescence, Photosynthetic Active Radiation (PAR), Direct Photosynthetic Active Radiation (SPAR), turbidity, and salinity.

5.2 Sequencing and processing

All the samples from Malaspina (116) were filtered in two size fractions: small (0.2 μm) and large (3 μm) using 142 mm polycarbonate membrane filters. The small fraction comprised free-living bacteria, archaea, and picoeukaryotes, while the large fraction contained larger eukaryotes and particle-attached microorganisms. This study focused solely on samples from the small fraction. DNA sequencing was conducted using Illumina MiSeq with a 2 x 250 bp reads configuration, yielding a total read count ranging from 319 to 508 million reads per sample. Initial read processing involved contaminant scanning (e.g., Illumina adapter sequences) and quality-based cleaning using FastQC [49], FastX-Toolkit [50], and SolexaQA [51] software. Trimmed reads falling below 45 nucleotides and quality score 33 were subsequently discarded. Functional annotation was done by homology search with DIAMOND v.0.9.13 [52] against

the NCBI database for taxonomic assignment and KO (KEGG Orthology) and COG databases for automatic functional assignment. All relevant steps were conducted with SqueezeMeta v.1.5.1 [53], including comprehensive analysis steps, multi-metagenome support for coassembling related metagenomes and retrieving individual genomes through binning procedures.

5.3 Functional delineation of metagenomic queries related to nitrogen cycling

To explore nitrogen cycling, we reconstructed reference phylogenetic trees for various marker genes as per Rivas-Santisteban *et al.* [12]. These are available in File S1. We constructed reliable trees to discern true markers from similar paralogues. From metagenomic queries, we identify candidate genes using specific protein models (PFAM, COGs, KEGG). These genes are then classified within our high-quality curated tree. We established reference databases for 18 genes encoding proteins involved in nitrogen cycling (Table S1), covering N redox functions (*amoA*, *nifH*, *nirK*, *ureC*, *uca*, *nxrA*, *nirS*, *nasA*, *narG*, *nosZ*, *hzsA*), nitrogen transporters (*amt*, *nrtA*, *tauA*, *potF*, *livJ*, *bztA*), and biosynthetic pathways (*ilvC*). After tree construction, candidate genes from all retrieved metagenomic queries were identified based on SqueezeMeta predictions using the provided specific protein models. These sequences were placed in the reference phylogenetic trees to distinguish particular sequences with biochemical evidence. We tagged the trees and placements with MetaTag v.0.1.1; [54]. True marker genes were retained, and their coverage and abundance were estimated using the SqueezeMeta pipeline.

5.4 Assessment of protein diversification

We assessed the expected sequence richness by fitting

$$d_{exp,s} = c_s A_s^{\gamma_s} \tag{1}$$

where s were sequences considered for a taxon, implementation and environment, abundance was measured in transcript per million (TPM), richness was the natural number of unique sequences, and where c_s and γ_s were constants. Thus, *delta* was calculated as a coefficient between $d_{obs,s}$ and $d_{exp,s}$. Kernel density estimates of Figures S3 & S4 were obtained with custom software. More details can be found in Figure S6.

5.5 Generalized Linear Model

This study employed a generalized linear model (GLM) to analyze the relationship between the predictor variables (contexts) and the response variable (*delta*). A GLM with a Gaussian distribution and an identity link function was specified for the analysis. This model was chosen based on the continuous nature of the response variable and the assumption of a linear relationship between the predictor and response variables.

The GLM was fitted to the data using the iterative reweighted least squares (IRLS) method. This method iteratively estimates the model parameters to minimize the

residual sum of squares. The fit of the GLM was assessed using several metrics, including the log-likelihood, deviance, Pearson chi-square, and pseudo R^2 (Cragg-Uhler). These metrics were used to evaluate how well the model captured the variation in the response variable and to assess the overall goodness-of-fit. All statistical analyses were conducted with a significance level of $\alpha = 0.05$.

Supplementary information. More information can be found in the attached supplementary file.

Acknowledgments. We thank P. Yubero for discussion.

Declarations

- Funding
  Funded by project TRAITS (PID2019-110011RB-C31/-C32) of Agencia Estatal de Investigación, Spanish National Plan for Scientific and Technical Research and Innovation. Funded by MCIN/AEI/10.13039/501100011033 from the Spanish Ministerio de Ciencia e Innovación and the European Union ("NextGenerationEU"/PRTR). Work supported by Ph.D. fellowship PRE2020-096130 from the Spanish Ministerio de Ciencia e Innovación and the European Social Fund.
- Conflict of interest/Competing interests
  The authors declare no competing interests.
- Ethics approval
  Not applicable.
- Consent to participate
  Not applicable.
- Consent for publication
  The authors consent to publication.
- Availability of data and materials
  Request for data and materials should be addressed to J.R.S.
- Code availability
  Code for guild display is available in https://github.com/pyubero/microguilds. Custom code for diversification kernel density estimations on request.
- Authors' contributions
  J.R.S.: Conceptualization, Methodology, Validation, Writing - Original Draft, Investigation, Data curation, Visualization, Software, Formal analysis; N.F.G.: Methodology, Software, Writing - Review & Editing; R.L.P.: Methodology, Software, Writing - Review & Editing; J.T. and C.P.A.: Conceptualization, Writing - Review & Editing, Project administration, Supervision, Funding acquisition.

References

[1] Vernadsky, V.I.: The Biosphere, pp. 121–125. Springer, Berlin (1998)

[2] Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M., Sauer, U.: Multidimensional optimality of microbial metabolism. Science **336**(6081), 601–604 (2012)

[3] Fuhrman, J.A., Cram, J.A., Needham, D.M.: Marine microbial community dynamics and their ecological interpretation. Nature Reviews Microbiology **13**(3), 133–146 (2015)

[4] Zhou, Z., Tran, P.Q., Kieft, K., Anantharaman, K.: Genome diversification in globally distributed novel marine proteobacteria is linked to environmental adaptation. The ISME Journal **14**(8), 2060–2077 (2020)

[5] Falkowski, P.G.: The role of phytoplankton photosynthesis in global biogeochemical cycles. Photosynthesis research **39**, 235–258 (1994)

[6] Arrigo, K.R.: Marine microorganisms and global nutrient cycles. Nature **437**(7057), 349–355 (2005)

[7] Matear, R.J., Wang, Y.-P., Lenton, A.: Land and ocean nutrient and carbon cycle interactions. Current Opinion in Environmental Sustainability **2**(4), 258–263 (2010)

[8] Crowe, S.A., Treusch, A.H., Forth, M., Li, J., Magen, C., Canfield, D.E., Thamdrup, B., Katsev, S.: Novel anammox bacteria and nitrogen loss from lake superior. Scientific reports **7**(1), 13757 (2017)

[9] Engström, P., Dalsgaard, T., Hulth, S., Aller, R.C.: Anaerobic ammonium oxidation by nitrite (anammox): implications for n2 production in coastal marine sediments. Geochimica et Cosmochimica Acta **69**(8), 2057–2065 (2005)

[10] Kuenen, J.G.: Anammox bacteria: from discovery to application. Nature Reviews Microbiology **6**(4), 320–326 (2008)

[11] Fay, A., McKinley, G.: Global open-ocean biomes: mean and temporal variability. Earth System Science Data **6**(2), 273–284 (2014)

[12] Rivas-Santisteban, J., Yubero, P., Robaina-Estevez, S., Gonzalez, J.M., Tamames, J., Pedros-Alio, C.: Quantifying microbial guilds. bioRxiv (2023) https://doi.org/10.1101/2023.07.23.550202

[13] Duarte, C.M.: Seafaring in the 21st century: the malaspina 2010 circumnavigation expedition (2015)

[14] Acinas, S.G., Sánchez, P., Salazar, G., Cornejo-Castillo, F.M., Sebastián, M., Logares, R., Royo-Llonch, M., Paoli, L., Sunagawa, S., Hingamp, P., *et al.*: Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. Communications Biology **4**(1), 604 (2021)

[15] Karner, M.B., DeLong, E.F., Karl, D.M.: Archaeal dominance in the mesopelagic zone of the pacific ocean. Nature **409**(6819), 507–510 (2001)

[16] Kampouris, I.D., Gründger, F., Christensen, J.H., Greer, C.W., Kjeldsen, K.U., Boone, W., Meire, L., Rysgaard, S., Vergeynst, L.: Long-term patterns of hydrocarbon biodegradation and bacterial community composition in epipelagic and mesopelagic zones of an arctic fjord. Journal of hazardous materials **446**, 130656 (2023)

[17] Longhurst, A.R.: Ecological Geography of the Sea. Elsevier, ??? (2010)

[18] Capone, D.G.: The marine nitrogen cycle. Microbe Magazine **3**(4), 186–192 (2008)

[19] Sharp, J.H.: The distributions of inorganic nitrogen and dissolved and particulate organic nitrogen in the sea. Nitrogen in the marine environment, 1–35 (1983)

[20] Kirchman, D.L., Gasol, J.M.: Microbial Ecology of the Oceans, pp. 361–364. John Wiley & Sons, ??? (2018)

[21] Ward, B.B.: Molecular approaches to marine microbial ecology and the marine nitrogen cycle. Annu. Rev. Earth Planet. Sci. **33**, 301–333 (2005)

[22] Casciotti, K.L.: Nitrite isotopes as tracers of marine n cycle processes. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **374**(2081), 20150295 (2016)

[23] Hutchins, D.A., Capone, D.G.: The marine nitrogen cycle: new developments and global change. Nature Reviews Microbiology **20**(7), 401–414 (2022)

[24] Bronk, D.A., Glibert, P.M., Ward, B.B.: Nitrogen uptake, dissolved organic nitrogen release, and new production. Science **265**(5180), 1843–1846 (1994)

[25] Kuypers, M.M., Marchant, H.K., Kartal, B.: The microbial nitrogen-cycling network. Nature Reviews Microbiology **16**(5), 263–276 (2018)

[26] Weber, T.S., Deutsch, C.: Ocean nutrient ratios governed by plankton biogeography. Nature **467**(7315), 550–554 (2010)

[27] Kobayashi, S., Hira, D., Yoshida, K., Toyofuku, M., Shida, Y., Ogasawara, W., Yamaguchi, T., Araki, N., Oshiki, M.: Nitric oxide production from nitrite reduction and hydroxylamine oxidation by copper-containing dissimilatory nitrite reductase (nirk) from the aerobic ammonia-oxidizing archaeon, nitrososphaera viennensis. Microbes and environments **33**(4), 428–434 (2018)

[28] Deng, C., Cheng, C.-H.C., Ye, H., He, X., Chen, L.: Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. Proceedings of the National Academy of Sciences **107**(50), 21593–21598 (2010)

[29] Kalvelage, T., Jensen, M.M., Contreras, S., Revsbech, N.P., Lam, P., Günter, M., LaRoche, J., Lavik, G., Kuypers, M.M.: Oxygen sensitivity of anammox and coupled n-cycle processes in oxygen minimum zones. PloS one **6**(12), 29299 (2011)

[30] Van Teeseling, M.C., Neumann, S., Van Niftrik, L.: The anammoxosome organelle is crucial for the energy metabolism of anaerobic ammonium oxidizing bacteria. Microbial Physiology **23**(1-2), 104–117 (2013)

[31] Arístegui, J., Gasol, J.M., Duarte, C.M., Herndld, G.J.: Microbial oceanography of the dark ocean's pelagic realm. Limnology and Oceanography **54**(5), 1501–1529 (2009)

[32] Bowler, C., Karl, D.M., Colwell, R.R.: Microbial oceanography in a sea of opportunity. Nature **459**(7244), 180–184 (2009)

[33] Grossart, H.-P., Massana, R., McMahon, K.D., Walsh, D.A.: Linking metagenomics to aquatic microbial ecology and biogeochemical cycles. Limnology and Oceanography **65**, 2–20 (2020)

[34] Bouman, H.A., Ulloa, O., Barlow, R., Li, W.K., Platt, T., Zwirglmaier, K., Scanlan, D.J., Sathyendranath, S.: Water-column stratification governs the community structure of subtropical marine picophytoplankton. Environmental Microbiology Reports **3**(4), 473–482 (2011)

[35] Shi, Y., Tyson, G.W., Eppley, J.M., DeLong, E.F.: Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the open ocean. The ISME journal **5**(6), 999–1013 (2011)

[36] Alcolombri, U., Peaudecerf, F.J., Fernandez, V.I., Behrendt, L., Lee, K.S., Stocker, R.: Sinking enhances the degradation of organic particles by marine bacteria. Nature Geoscience **14**(10), 775–780 (2021)

[37] Cael, B., Cavan, E.L., Britten, G.L.: Reconciling the size-dependence of marine particle sinking speed. Geophysical Research Letters **48**(5), 2020–091771 (2021)

[38] Ciccarese, D., Tantawi, O., Zhang, I.H., Plata, D., Babbin, A.R.: Microscale dynamics promote segregated denitrification in diatom aggregates sinking slowly in bulk oxygenated seawater. Communications Earth & Environment **4**(1), 275 (2023)

[39] Zehr, J.P., Capone, D.G.: Unsolved mysteries in marine nitrogen fixation. Trends in Microbiology (2023)

[40] Cullen, J.J.: Hypotheses to explain high-nutrient conditions in the open sea. Limnology and Oceanography **36**(8), 1578–1599 (1991)

[41] Moore, C., Mills, M., Arrigo, K., Berman-Frank, I., Bopp, L., Boyd, P., Galbraith,

E., Geider, R., Guieu, C., Jaccard, S., *et al.*: Processes and patterns of oceanic nutrient limitation. Nature geoscience **6**(9), 701–710 (2013)

[42] Medina-Silva, R., Oliveira, R.R., Pivel, M.A., Borges, L.G., Simão, T.L., Pereira, L.M., Trindade, F.J., Augustin, A.H., Valdez, F.P., Eizirik, E., *et al.*: Microbial diversity from chlorophyll maximum, oxygen minimum and bottom zones in the southwestern atlantic ocean. Journal of Marine Systems **178**, 52–61 (2018)

[43] Hurst, L.D.: The ka/ks ratio: diagnosing the form of sequence evolution. TRENDS in Genetics **18**(9), 486–487 (2002)

[44] Kreitman, M.: The neutral theory is dead. long live the neutral theory. Bioessays **18**(8), 678–683 (1996)

[45] Cohan, F.M., Koeppel, A.F.: The origins of ecological diversity in prokaryotes. Current Biology **18**(21), 1024–1034 (2008)

[46] Merbt, S.N., Stahl, D.A., Casamayor, E.O., Martí, E., Nicol, G.W., Prosser, J.I.: Differential photoinhibition of bacterial and archaeal ammonia oxidation. FEMS microbiology letters **327**(1), 41–46 (2012)

[47] Newell, S.E., Babbin, A.R., Jayakumar, A., Ward, B.B.: Ammonia oxidation rates and nitrification in the arabian sea. Global Biogeochemical Cycles **25**(4) (2011)

[48] Margalef, R.: Diversity, stability and maturity in natural ecosystems. In: Unifying Concepts in Ecology: Report of the Plenary Sessions of the First International Congress of Ecology, The Hague, the Netherlands, September 8–14, 1974, pp. 151–160 (1975). Springer

[49] Andrews, S.: Fastqc version 0.11. 5. A Quality Control Tool for High Throughput Sequence Data (2016)

[50] Gordon, A., Hannon, G.: FASTX-Toolkit. FASTQ/A short-reads pre-processing tools (version 0.0. 13) (2010)

[51] Cox, M.P., Peterson, D.A., Biggs, P.J.: Solexaqa: At-a-glance quality assessment of illumina second-generation sequencing data. BMC bioinformatics **11**(1), 1–6 (2010)

[52] Buchfink, B., Xie, C., Huson, D.H.: Fast and sensitive protein alignment using diamond. Nature methods **12**(1), 59–60 (2015)

[53] Tamames, J., Puente-Sánchez, F.: Squeezemeta, a highly portable, fully automatic metagenomic analysis pipeline. Frontiers in microbiology **9**, 3349 (2019)

[54] Robaina Estévez, S., Fernández González, N., González Hernández, J.M.: metaTag: Functional and Taxonomical Annotation of Metagenomes Through Phylogenetic Tree Placement. [Computer software]. https://doi.org/10.5281/

zenodo.7949949