Picoplankton nitrogen guilds in the tropical and subtropical oceans: from the surface to the deep

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, Pozuelo de Alarcón, Spain.
[3]Biogeochemistry and Microbial Ecology Department, Museo Nacional de Ciencias Naturales (MNCN-CSIC), Madrid, Spain.

*Corresponding author(s). E-mail(s): jtamames@cnb.csic.es; cpedros@cnb.csic.es;
†These authors contributed equally to this work.

Abstract

Ecological guilds quantify the incidence and extent of resource transformation functions, irrespective of the species involved. Therefore, tackling the microbial nitrogen guilds is key to our understanding of the oceanic nitrogen cycle, but quantitative estimates of guild contribution across varying depths and under specific environmental conditions have yet to be accomplished. In this study, we examine the main picoplankton guilds participating in nitrogen cycling 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 site specific conditions. For example, differential conditions such as nitrogen depletion and oxygen availability affected the two groups of guilds in different ways. These findings have implications for the understanding of global nitrogen fluxes and the biosphere's functional diversification.

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

1 Introduction

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, irradiance, or hydrostatic pressure [1–7]. Under such ever-changing conditions, how can we trace causality in the observed microbial functional responses?

Microbial functional response to these variables have been largely explored [8–12], although, as to our knowledge, they have not yet been depicted from a quantitative framework such as ecological guilds. An ecological guild refers to a list of organisms exploiting the *same resource in the same way* across environments, regardless of their phylogenetic relatedness [13]. Membership is not a static quality, and the contribution of a member needs to be quantified in each context. This framework can thus disentangle the contributions of predictor variables to the functional profile. To this end, we screen the functional responses to differential environments in two ways: (1) *how* the function has been carried out (incidence and number of unique protein variants, and how each of them acts on the substrate/s), and (2) *who* implemented the necessary proteins (taxon-aware effects).

In the present contribution our purpose was to analyze the microbial guilds involved in the N cycle in the oceans with the guild approach. The N cycle is essential to Earth's biogeochemistry, transforming inert N into forms that sustain life, regulate ecosystems, and impact climate. Biotic processes play a fundamental role in the allocation and rate of the N cycle [14]. The reasons for understanding microbial performance regarding N cycling are manifold: (i) there exist up to seven redox states in the range of possible N compounds in the ocean, all of which may be differently transformed by microbes [10, 15]; (ii) they consist of both organic and inorganic N species [16]; (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 [17]; (iv) there is variability in N-limitation among different areas; and (v) N compounds are differentially limiting, for example, N_2 is never limiting, whereas nitrate (NO_3-) frequently limits phytoplankton growth [15, 17]. Thus, the oceanic N cycle depends on numerous actors yet, in turn, its proper functioning is critical for the maintenance of ecosystems.

We quantified microbial nitrogen cycle guilds within low and mid-latitude oceanic biomes using metagenomic data from the Malaspina 2010 (MP) expedition [9, 18]. Sampling was done with Conductivity-Temperature-Depth (CTD) profilers with Niskin bottles from the surface to 4,000 meters, analyzing them for physical, chemical, and biological parameters. We classified the MP metagenomes according to their depth, establishing three broad categories: epipelagic (0-200 m depth), mesopelagic (200-1000 m depth) and bathypelagic (1,000-4,000 m depth). This depth stratification is commonly used to study marine ecology, as these defined oceanic layers often exhibit similar physicochemical characteristics [19, 20]. However, geographic location is also crucial in defining equivalent oceanic physicochemical regions. Longhurst's provinces represent one effort to systematize these zones [21]. Therefore, we considered both depth and specific locations to study how they influence the abundance and diversification of oceanic microbial functions (Fig. 1a).

Akin to our categorization of the MP samples using depth-location criteria, we also classified the guilds related to the N cycle according to the 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 the first, it comprises transporters for different N compounds. Within the second, it includes key proteins for the transformation processes where the imported N compounds act as electron donors or acceptors. With this clear division between biomes and guilds, we described their interplay as inferred from the MP metagenomes.

2 Results & Discussion

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 considered three depth intervals separately. We calculated N/P ratios to identify stations departing from the ocean average. This ratio provides a rough idea of how limiting is the dissolved N in an environment [22]. We assumed the ratios were causally linked to common nitrogen-related genes in various ocean regions. When we calculated this ratio in MP samples (Fig. 1b), the values showed less variability below mesopelagic depths, where they converged, being only marginally lower than the Redfield ratio (16:1), which is considered not to be either N or P limiting [23]. In effect, even though N/P vertical profiles were very similar in most stations, there were two that showed significant differences and we looked at them separately (Fig. 1b). The first was a water column off the coast of Mexico (corresponding to station 120), where an Oxygen Minimum Zone (OMZ) occurs. As expected, this station had a very low mesopelagic maximum (N/P $\approx$ 12.4), and its values were always lower than the samples from other stations below 600m, as shown in Figure 1b. The second was a station from the open North Atlantic Ocean (corresponding to station 141), whose mesopelagic maximum was 18.96 and values were always higher than the other stations below this point (430m).

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. We decided to analyze these stations separately. Finally, we also chose to analyze the SAtl station because, intriguingly, we detected anammox genes in an aerobic environment. Thus, we will show results for the average ocean (all other stations) and for each one of these three stations separately.

2.2 Taxonomic and depth distribution of N functional markers

We assembled the metagenomes, predicted the proteins, mapped the reads, and assigned the taxonomy for each marker gene (Table S1) as described in Methods. Our

annotation includes a step to purge false positives of the target function by using reference trees (for details see Rivas-Santisteban *et al.* [13]). Briefly, we build a tree with available sequences of a given gene and place the metagenomic reads in this tree. An example is shown in Figure S1 for the *hzsA* gene, a step in the anammox process. The tree has several different clusters. In particular the *hzsA-I* cluster includes all the know sequences belonging to *Planctomycetes* and thus, genes within this cluster are expected to carry out the function. On the other hand, sequences placed in cluster *hzsA-III* do not belong to any known anammox microorganisms and, further, we found them in an aerobic environment (SAtl station). Therefore, we discarded the sequences in cluster *hzsA-III*. We carried out the same procedure for every gene. The amounts of discarded sequences for nine genes are shown in Figure S2. Some genes had almost no discarded sequences, for example *amt*. The *livJ* gene, on the other hand, had many discarded sequences. These differences depend on how well characterized is each gene and how similar are sequences of other genes that can be considered as paralogs.

The abundance of the N functional markers (in normalized *features per million*, FPM) in the 75 samples appear in Figure 2). The taxonomic distribution for relative FPM *per gene per layer* is shown in Fig. 2a–c, so that the composition of the less abundant markers can be appreciated. The cumulative FPM is represented in Figure 2d–e. The mean and standard deviation for the abundance of gene ORFs within each layer are shown in Table S2. The most abundant genes were *amt* (marker for ammonia transport), *ilvC* (marker for branched amino acid synthesis), *nirK* (marker for denitrification or potential nitrification in archaea [24]), and *ureC* (marker for urea hydrolysis) regardless of depth. There was considerable species turnover between the epipelagic and mesopelagic, but the latter and the bathypelagic were similar. Primary producers (*Cyanobacteria*) and *Alphaproteobacteria* dominated at surface waters, while they were replaced by *Nitrososphaeria* with depth. As a consequence, the *nirK* gene became more abundant. The marker for N-fixation (*nifH*) was found in extremely few counts and it did not appear in the mesopelagic (Figure 2b & e). This is because most of the counts were filtered out as belonging to cluster IV of the *nifH* tree, which is widely considered non-functional [25] (Fig. S2). Finally, some functional markers were found profusely only at certain depths. For example, nitrate reduction (*nasA*) was three times more abundant in the bathypelagic than at other depths.

Overall, the most remarkable features of this distribution were: (1) that the most frequent genetic implementation for N uptake function in the oceans was the ammonium transporter; (2) that the metabolic capacity to build branched chain amino acids was very extended across oceanic layers and taxa (ubiquity of *ilvC*); and (3) that the main N redox process in the oceans involved *nirK* functions (denitrification/nitrification partial steps) only below the euphotic zone. These genes were mainly implemented by *Nitrososphaeria*. In the next section we describe in detail spatial and compositional trends of the main guilds.

2.3 Description of oceanic N guilds

We quantified guilds using k-values (see methods). A full discussion of this approach can be found in Rivas-Santisteban et al. (2024) [13]. Briefly, these k-values include an estimate of the abundance of a given gene (A, number of features or reads of the gene)

and an estimate of the unexpected diversity of sequences within such a gene (delta value). Generally, the diversity of sequences of a gene increases with its abundance. We show this relationship for both N uptake and N redox genes in Figure S3. As shown, a relationship can be calculated for each gene from the environmental data providing the expected diversity (d_{exp}). Then, the actual diversity for a given gene in a given environment (d_{obs}) can be divided by the expected diversity. Thus, genes with an observed diversity equal to that expected from their abundance will have a delta value of 1. Genes with a diversity of sequences larger than that expected from their abundance are rewarded with a higher value of delta, while those with a lower diversity are penalized. Then the delta value is multiplied times the abundance to obtain the k value. We do this because those genes with a higher diversity are expected to confer a higher stability and versatility to the guild under changing conditions. For a clear presentation, k-values were log transformed and displayed as radial plots (Figs. 3 & 4).

2.3.1 N uptake guilds

The N uptake genes could be divided between those incorporating inorganic N and those incorporating organic N. The marine microbiotas showed a general pattern in the relative abundances of these genes (Fig. 3). Among the former, ammonia uptake (*amt*) was the most popular function, while the nitrate uptake gene (*nrt*) was always present but in much lower abundance. Finally, the nitrogen fixation gene (*nifH*) was almost below detection levels in all stations. This is in accordance with the energy required to reduce each compound to the redox level of organic matter: ammonia is already reduced, nitrate needs reduction to ammonia and nitrogen gas requires the very expensive N fixation process. The nitrate uptake gene was present at all depths but its abundance increased in the bathypelagic. The only exception to this pattern was in the mesopelagic of the SAtl station where *nrtA* found its largest abundance.

We had no previous hypothesis as to which of the organic-N uptake mechanisms would be more important. There are plenty of N-compounds in the oceans. We chose three genes responsible for the uptake of amino acids and three for other N-compounds. Among the former, the gene for the uptake of carboxylated amino acids *bztA* was the most abundant at all depths and stations. The other two genes (*livJ* for branched-chain amino acids and *aapJ*, a very general amino acid transporter) did change in abundance from one sample to another but they were always found in very low abundance compared to *bztA*.

Among the transporters of other N-compounds *ssuA* (transporter of sulfonates that have both nitrogen and sufur) was always very rare. The other two, *potF* for the transport of polyamines and *tauA* transporting taurine were fairly abundant in general but with differences. Gene *tauA* was more abundant than *potF* in the epi- and mesopelagic, while *potF* was more abundant in the bathypelagic.

These patterns were very robust since they appeared in almost all stations at all depths and, thus, they reflect the relative importance of the different N-compounds in the ecology of marine microbial communities. It is true that presence of a gene does not imply activity. However, the presence or absence of certain genes and not others is a strong indicator of their eco-evolutionary relevance [26, 27].

2.3.2 N-redox transformation guilds

Nitrogen redox guilds were much more variable among depths and stations than nitrogen uptake guilds (Fig. 4).

The potential for nitrification was present in most stations in the meso- and bathypelagic. Two taxa assigned to *Nitrosopelagicus* were the main contributors. There were two exceptions to this pattern. Neither station SAtl nor OMZ showed presence of the *amoA* gene. In the mesopelagic of the OMZ station oxygen was scarce and this is needed for nitrification. At the other depth intervals, however, there was plenty of oxygen. In OMZ station, the N/P ratio was the lowest found and, perhaps, this relative scarcity of nitrogen prevented nitrification from being an important process. The SAtl station, however, would require a different explanation. We will discuss this station below.

The OMZ station showed a very different pattern from that of the average ocean. Particularly in the mesopelagic, the depth of the oxygen minimum zone, had large abundances of both anammox and denitrification genes. This is reasonable since both processes take place in the absence, or at very low concentrations, of oxygen. The denitrification gene *narG* was present in all stations and depths, and its abundance generally increased from the epi- to the deeper sections. However, it was much more abundant in the oxygen minimum zone, as well as the other two genes examined (*napA* and *nosZ*).

Anammox genes were only found in the OMZ and, intriguingly, in the SAtl station. Thanks to our protocol to place metagenomic sequences in a phylogenetic tree of the available gene sequences, we found that the sequences from OMZ were placed in a clade with Planctomycetes-like *hzsA* sequences. Those from the SAtl station, on the other hand, appeared in another clade with no known anammox sequences (Figure S1). The latter were assigned to two *Alcanivorax* taxa. As mentioned, these waters were fully oxygenated and, thus, anammox was not expected. *Alcanivorax* is not known to carry out anammox or to have anammoxosomes. This gene (labeled *hzsA*-III) is a paralog that has probably acquired a different function. We do not have data to explain the proliferation of *Alcanivorax* at station SAtl, but it certainly changed the distribution of the nitrogen redox genes in these waters. Our *hzsA* metagenomic sequences were mostly assigned to *Planctomycetes*. However, there were several that were closer to those of *Verrucomicrobia*. The *hzsA*-I cluster built with data base sequences (Figure S1) was mostly composed of *Planctomycetes*, but it also had sequences from *Verrucomicrobia* and even one assigned to a *Clostridium*, despite the fact that these bacteria have not been shown to carry our anammox or to have anammoxosomes. These assignments are likely the consequence of a poor characterization of the diversity of *Planctomycetes* sequences in nature. Thus, we conclude that these sequences likely belonged to not well characterized *Planctomycetes*.

2.4 Diversification of N cycle proteins

One important quantitative aspect of a guild is certainly its abundance. If a given guild is twice as abundant in one sample as in another, this strongly suggest it is more important in the former sample. However, there is a second aspect. A guild could be

represented by many identical sequences. For example, this could happen if there is a bloom of one taxon possessing the gene. On the other hand, the guild may be formed by many different sequences (even though able to carry out the function, naturally) irrespective of the total abundance. We postulate that this diversity of sequences within a guild is a relevant parameter since presumably higher diversity of sequences will provide stability, adaptability and resilience to the guild [28]. Our quantification method allows separating these two factors and to analyze them independently.

The density functions for delta values are shown in Figure 5. For any given sample (or collection of samples, for example the epipelagic samples of the OMZ station). The delta values of each taxon will be distributed along the x axis. When the observed diversity of proteins is equal to the expected diversity, delta will have a value of 1, and the corresponding data point will appear in the x-axis with a value of 0 (since the scale is logarithmic). If the observed diversity is higher than expected, the delta will have a positive value (displaced to the right of the 0) and when the observed diversity is lower than expected it will appear to the left. The accumulation of real points at each x value will result in a density function that approximates the distribution of delta values in the sample. The statistical method used to find and smooth this density function is the kernel density estimation. The result allows a visual comparison of the distribution of delta values among different samples (or collections of samples).

Let's look at the OMZ station first. In the case of the N uptake guilds the density functions for epi- and bathypelagic are fairly similar and centered around the value of delta = 1. This distribution is the most likely expected one. The curve for the mesopelagic, however, shows a displacement towards the left, indicating that many of the taxa showed a diversity of proteins lower than expected. On the other hand, this did not happen in the case of the N redox guilds.

In most cases, the curves were centered at the value of delta = 1, but with a certain dispersion of values towards the left: all the stations for N redox genes and NAtl and OMZ, SAtl for N uptake genes (but not the average ocean). The only exception to this distribution was the epipelagic from the SAtl station, that showed a density function displaced towards the right, indicating a diversity of proteins larger than expected from their abundance.

The general distribution, then, shows that most marine samples have most of their taxa with the diversity expected from their abundance, but in many cases, there is a subset of taxa with diversities lower than expected. In some particular cases, there is a clear departure from this general distribution. In the case of the OMZ bathypelagic, the N uptake genes showed lower diversity that expected. One possibility is that the absence of oxygen decreases the number of taxa that can live here. In the case of the epipelagic of the SAtl station the higher diversity may be due to the presence of oil degrading bacteria such as *Alcanivorax*. The area has been known to experience oil spills that would presumably provide a large variety of organic compounds derived from oil. Perhaps this requires a larger variety of proteins with slightly different affinities.

The equivalent graphs considering the abundance of the functional genes only (and not the diversities) are shown in Figure S4. In this case there were very few differences between N uptake and N redox genes. Also, the secondary maximum in

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

	N_{uptake}			N_{redox}		
Context	coef	std. err	$P > \|z\|$	coef	std. err	$P > \|z\|$
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 *

most distributions was found to the right, not to the left as in the diversity graphs. And, finally, the main deviation from the general pattern was found again in the mesopelagic of the OMZ station. In this case, however, the secondary maximum was displaced towards larger abundances.

It is apparent than considering abundance alone misses very interesting patterns. Having a diversity larger than expected from abundance suggests a guild with larger capacity for adaptation to different circumstances and, thus presumably, more stable in time and space. And this is independent of the abundance of the sequences in that guild at any given point.

In order to further quantify these relationships, we applied generalized linear models. The coefficients and their significance appear in Table 1. These models allow a statistical comparison between one set of data (for example the samples from the OMZ station) to the rest of MP dataset. The resultant coefficient provides an indication of the strength of the difference as well as the significance of such difference if it exists. As shown in Table 1, the NAtl and SAtl stations did not differ significantly from the general ocean, neither in N uptake nor in N redox genes, while the OMZ station showed a rather significant negative relationship. The visual impression from Figure 4 is confirmed by the analysis. When looking at the three depth layers, the epipelagic did not show significant differences with the general ocean, but the other two layers did depart both in N uptake and N redox genes, in this case the departure was positive.

2.5 The guild approach

The guild concept is particularly appropriate to analyze metagenomic data with functional and taxonomic information. A guild is composed of organisms that carry out the same function, with independence of being taxonomically related or not. The guild is responsible for the function presumably with different compositions in different environments. For example, urea degradation, as represented by the gene *ureC*, had the same relative importance at the three depth intervals studied (Figure 2d, e and f) but the organisms with this gene were very different at the epipelagic (*Cyanobacteria* and *Alphaproteobacateria*) from those at depth (mostly *Nistrosophaeria*).

This pattern was not only found with depth, but also among stations with different environmental conditions. An example of this is the guild represented by the *napA* gene, that codes the molybdoprotein subunit of the periplasmic nitrate reductase, involved in denitrification. In the OMZ station (Figure 4, upper row) the main contributor to this gene in the oxygen minimum zone (mesopelagic of station OMZ) was a *Desulfovibrio* strain, but it was absent from both the bathy- and epipelagic

zones of this same station, where the guild was less important but still present. On the other hand, *Cupriavidus* was an important contributor to this guild in the epi- and the mesopelagic of the OMZ station. It was absent from the bathypelagic but it contributed again to this guild in the meso- and bathypelagic of the average ocean (Figure 4, lower row).

These examples show clearly the uncoupling between taxonomy and function that usually occurs in microbial communities. In other words, functions cannot be easily predicted from taxonomic composition alone. Our quantification and graphical presentation allow an easy analysis of the two types of information together.

3 Conclusions

Our study has some limitations. First, MP expedition only sampled tropical and subtropical waters but covered a substantial part of the Atlantic, Indian and Pacific Oceans. Thus, we cannot extrapolate to more temperate or polar waters. Next, the number of stations was also limited. However, the whole vertical profiles were analyzed down to 4000 m and our "average ocean" showed robust patterns. Finally, we only studied the picoplankton. Some functions, for example N fixation are known to be carried out in many cases by *Trichodesmium* aggregates or by symbionts of diatoms. We would have missed these genes. Despite this, diatoms are not generally abundant at the latitudes examined and our conclusions still hold for the picoplankton. Free-living diazotrophs are fairly rare in marine waters, accounting for up to 1% of the cells. Even though their function is obviously important [29–31], our considerations about N uptake preferences by picoplankton are still real and consistent with the energy required to incorporate the different N compounds into organic matter. In summary, we found fairly robust patterns in the importance and distribution of N guilds in the tropical and subtropical oceans.

Despite departures at particular samples, general patterns in the distribution of N guilds do exist in the ocean. Malaspina only sampled tropical and subtropical waters but covered a substantial part of the Atlantic, Indian and Pacific Oceans as well as a wide range of depths from the surface to 4000 m. These patterns included the following:

(i) The guilds for inorganic N uptake were in decreasing order of importance: ammonia, uptake, nitrate uptake, and nitrogen fixation throughout the ocean in accordance with the energy required to process each one of them.

(ii) Among the guilds for organic N uptake, the gene for carboxylated amino acids *bztA* was the most abundant at all depths and stations. In the case of non amino acid N organic compounds *potF* for the transport of polyamines and *tauA* transporting taurine were fairly abundant in general. Gene *tauA* was more abundant than *potF* in the epi- and mesopelagic, while *potF* was more abundant in the bathypelagic.

(iii) N uptake guilds were less variable in importance than N redox guilds.

(iv) In the OMZ area examined, anammox genes were dominated by *Planctomycetes*.

(v) *Alphaproteobacteria* and *Cyanobacteria* were the dominant contributors to the N guilds in the epipelagic, while *Nitrososphaeria* dominated the N cycle in the meso

and bathypelagic as shown by their contribution to N uptake (*amt*), denitrification (*nirK*) and nitrification (*amoA*), plus urea degradation (*ureC*).

(vi) Deviations from these general patterns have been summarized in Figure 6 for the OMZ and NAtl stations as compared to the average ocean.

(vii) Finally, considering not only abundance, but also unexpected diversity of sequences provides additional insight into the composition and variability of nitrogen guilds.

4 Methods

4.1 Sampling and sequencing

A description of the Malaspina cruise can be found in Duarte (2015) [18]. Raw reads, full sampling and sequencing methodology can be found for example in Villarino *et al.* (2022) [32] and Sánchez *et al.* [33]. These are briefly summarized here. 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. 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. The majority of samples originated from the Atlantic, Pacific, and notably the Indian Ocean, where many physicochemical parameters were documented for the first time during the expedition. Given the tropical and subtropical nature of the samples, seasonal climate variations are expected to have minimal impact. Metadata included date, latitude, longitude, depth, temperature, filter size, conductivity, oxygen levels, fluorescence, Photosynthetic Active Radiation (PAR), Direct Photosynthetic Active Radiation (SPAR), turbidity, and salinity.

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 [34], FastX-Toolkit [35], and SolexaQA [36] software. Trimmed reads falling below 45 nucleotides and quality score 33 were subsequently discarded.

4.2 Metagenomic workflow

All relevant first steps in the bioinformatic workflow were conducted with SqueezeMeta v.1.5.1 [37] predefined parameters. We assembled 75 metagenomes in sequential mode, retrieving individual genomes through binning procedures using DASTool software [38]. We limited the contig size to 200 bp with megahit [39]. Functional annotation was done by homology search with DIAMOND v.0.9.13 [40] against the NCBI database for taxonomic assignment and KO (KEGG Orthology) and COG databases for automatic functional assignment.

Abundance estimation was automatically calculated with SqueezeMeta v.1.5.1. We provided gene abundances as FPM or *features per million* (a neutral term for TPM that are not derived from a transcriptome), recalculated globally on the sum of raw reads from the 75 samples that are the subject of this study. Thus, the sum of all functional features across samples yield 1 million FPM.

4.3 Functional delineation of metagenomic queries related to nitrogen cycling

To explore nitrogen cycling, we established reference sequence databases for 18 selected genes encoding proteins strictly involved in key steps of the nitrogen cycling (Table S1), covering N redox functions (*amoA*, *nifH*, *nirK*, *ureC*, *uca*, *nxrA*, *nirS*, *nasA*, *narG*, *nosZ*, *hzsA*), nitrogen transporters – substrate-binding proteins (*amt*, *nrtA*, *tauA*, *potF*, *livJ*, *bztA*), and biosynthetic pathways (*ilvC*). These sequences include well-studied sequences with biochemical evidence, so that truly functional sequence spaces can be subsequently matched with short metagenomic queries. For this purpose, we reconstructed reference phylogenetic trees with the IQ-TREE software [41] for the above retrieved sequences, as per Rivas-Santisteban *et al.* [13]. These are available in File S1. We built these reliable trees as metagenomic sequence classifiers, to discern true markers from similar paralogs. The candidate genes were retrieved from metagenomes using the pertinent Hidden Markov Model [42] (PFAM, COGs, and/or KEGG) for each gene. Then, we identified truly-functional sequences based on the propagation of the functional annotations onto the reference trees. These candidate sequences were placed in the reference trees to distinguish particular sequences with biochemical evidence. We tagged the trees and placements with MetaTag v.0.1.1; [43]. True marker genes were retained for the functional analysis, and their coverage and abundance were estimated as depicted in 5.2.

4.4 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 *features per million* (FPM), 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 $d_{obs,s}$ (observed richness) and $d_{exp,s}$. Kernel density estimates of Figures 3 and S4 were obtained with custom software. Additional statistical tests and details can be found in Figure S6 and Table S3.

4.5 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 files.

Acknowledgments. We thank N. García-García for materials, P. Yubero and F. Puente-Sánchez for discussion. The last part of the work was funded by grant PID2023-146919NB-C21 from MICIU/AEI/10.13039/501100011033 and by FEDER, UE

Declarations

- Funding
  Funded by project 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
  Raw reads are deposited in the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/view/PRJEB52452) [33]. File S1 with the annotation reference trees and files is available on public repository (https://zenodo.org/records/14557548). Code for guild analysis is available in https://github.com/pyubero/microguilds.
- 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] Fay, A., McKinley, G.: Global open-ocean biomes: mean and temporal variability. Earth System Science Data **6**(2), 273–284 (2014)

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

[9] 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)

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

[11] Worden, A.Z., Cuvelier, M.L., Bartlett, D.H.: In-depth analyses of marine microbial community genomics. Trends in microbiology **14**(8), 331–336 (2006)

[12] Saunders, J.K., McIlvin, M.R., Dupont, C.L., Kaul, D., Moran, D.M., Horner, T., Laperriere, S.M., Webb, E.A., Bosak, T., Santoro, A.E., *et al.*: Microbial functional diversity across biogeochemical provinces in the central pacific ocean. Proceedings of the National Academy of Sciences **119**(37), 2200014119 (2022)

[13] Rivas-Santisteban, J., Yubero, P., Robaina-Estévez, S., González, J.M., Tamames, J., Pedrós-Alió, C.: Quantifying microbial guilds. ISME communications **4**(1), 042 (2024)

[14] Fowler, D., Coyle, M., Skiba, U., Sutton, M.A., Cape, J.N., Reis, S., Sheppard, L.J., Jenkins, A., Grizzetti, B., Galloway, J.N., *et al.*: The global nitrogen cycle in the twenty-first century. Philosophical Transactions of the Royal Society B: Biological Sciences **368**(1621), 20130164 (2013)

[15] Capone, D.G., Knapp, A.: The marine nitrogen cycle. MICROBE-AMERICAN SOCIETY FOR MICROBIOLOGY **3**(4), 186 (2008)

[16] 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)

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

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

[19] 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)

[20] 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)

[21] Longhurst, A.R.: Ecological geography of the sea, 80–85 (2010)

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

[23] Tett, P., Droop, M., Heaney, S.: The redfield ratio and phytoplankton growth rate. Journal of the Marine Biological Association of the United Kingdom **65**(2), 487–504 (1985)

[24] 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)

[25] Gaby, J.C., Buckley, D.H.: A global census of nitrogenase diversity. Environmental Microbiology **13**(7), 1790–1799 (2011)

[26] Diop, A., Torrance, E.L., Stott, C.M., Bobay, L.-M.: Gene flow and introgression are pervasive forces shaping the evolution of bacterial species. Genome Biology **23**(1), 239 (2022)

[27] Morris, J.J., Lenski, R.E., Zinser, E.R.: The black queen hypothesis: evolution of dependencies through adaptive gene loss. MBio **3**(2), 10–1128 (2012)

[28] 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

[29] Shao, Z., Xu, Y., Wang, H., Luo, W., Wang, L., Huang, Y., Agawin, N.S.R., Ahmed, A., Benavides, M., Bentzon-Tilia, M., *et al.*: Global oceanic diazotroph database version 2 and elevated estimate of global oceanic n 2 fixation. Earth System Science Data **15**(8), 3673–3709 (2023)

[30] Tang, W., Cerdán-García, E., Berthelot, H., Polyviou, D., Wang, S., Baylay, A., Whitby, H., Planquette, H., Mowlem, M., Robidart, J., *et al.*: New insights into the distributions of nitrogen fixation and diazotrophs revealed by high-resolution sensing and sampling methods. The ISME Journal **14**(10), 2514–2526 (2020)

[31] Pierella Karlusich, J.J., Pelletier, E., Lombard, F., Carsique, M., Dvorak, E., Colin, S., Picheral, M., Cornejo-Castillo, F.M., Acinas, S.G., Pepperkok, R., *et al.*: Global distribution patterns of marine nitrogen-fixers by imaging and molecular methods. Nature communications **12**(1), 4160 (2021)

[32] Villarino, E., Watson, J.R., Chust, G., Woodill, A.J., Klempay, B., Jonsson, B., Gasol, J.M., Logares, R., Massana, R., Giner, C.R., *et al.*: Global beta diversity patterns of microbial communities in the surface and deep ocean. Global Ecology and Biogeography **31**(11), 2323–2336 (2022)

[33] Sánchez, P., Coutinho, F.H., Sebastián, M., Pernice, M.C., Rodríguez-Martínez, R., Salazar, G., Cornejo-Castillo, F.M., Pesant, S., López-Alforja, X., López-García, E.M., *et al.*: Marine picoplankton metagenomes and mags from eleven vertical profiles obtained by the malaspina expedition. Scientific Data **11**(1), 154 (2024)

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

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

[36] 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)

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

[38] Sieber, C.M., Probst, A.J., Sharrar, A., Thomas, B.C., Hess, M., Tringe, S.G., Banfield, J.F.: Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nature microbiology **3**(7), 836–843 (2018)

[39] Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., Yamashita, H., Lam, T.-W.: Megahit v1. 0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods **102**, 3–11 (2016)

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

[41] Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A., Lanfear, R.: Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era. Molecular biology and evolution **37**(5), 1530–1534 (2020)

[42] Söding, J.: Protein homology detection by hmm–hmm comparison. Bioinformatics **21**(7), 951–960 (2005)

[43] 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

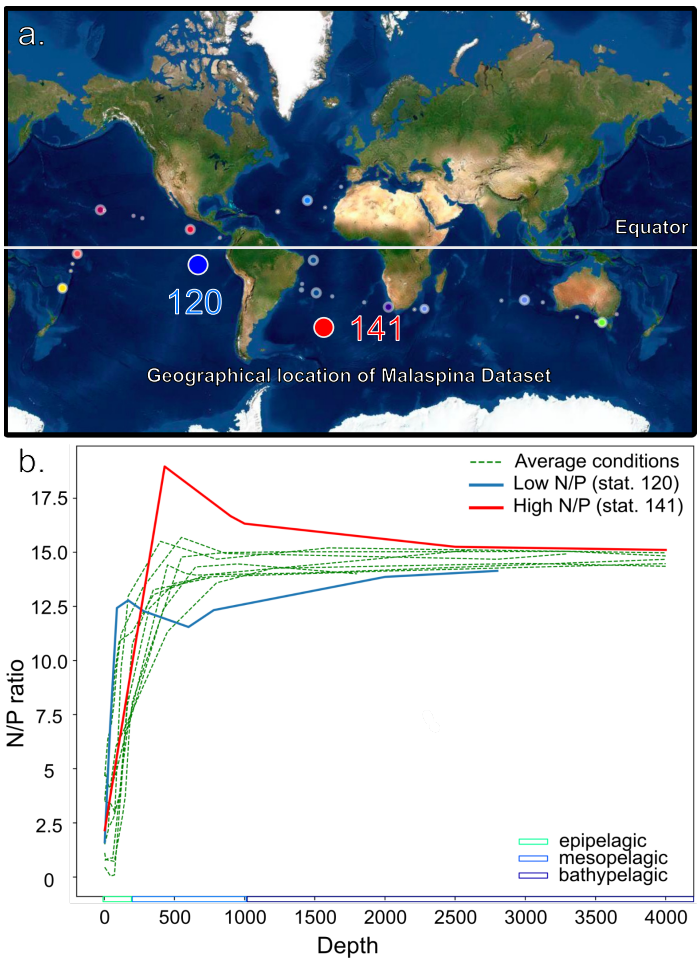

Fig. 1 a. Location of the samples and stations within the MP dataset. Bigger circles represent the stations with samples along all the vertical profile, and thus the ones analyzed in this contribution. b. Values of N/P ratios with depth in the 11 Malaspina water columns sampled. Ratios were very similar at the epi- and bathypelagic but two stations showed marked differences in the mesopelagic: station 120 (Pacific, off the Mexico coast, with an oxygen minimum zone) and station 141 (North Atlantic, open ocean).

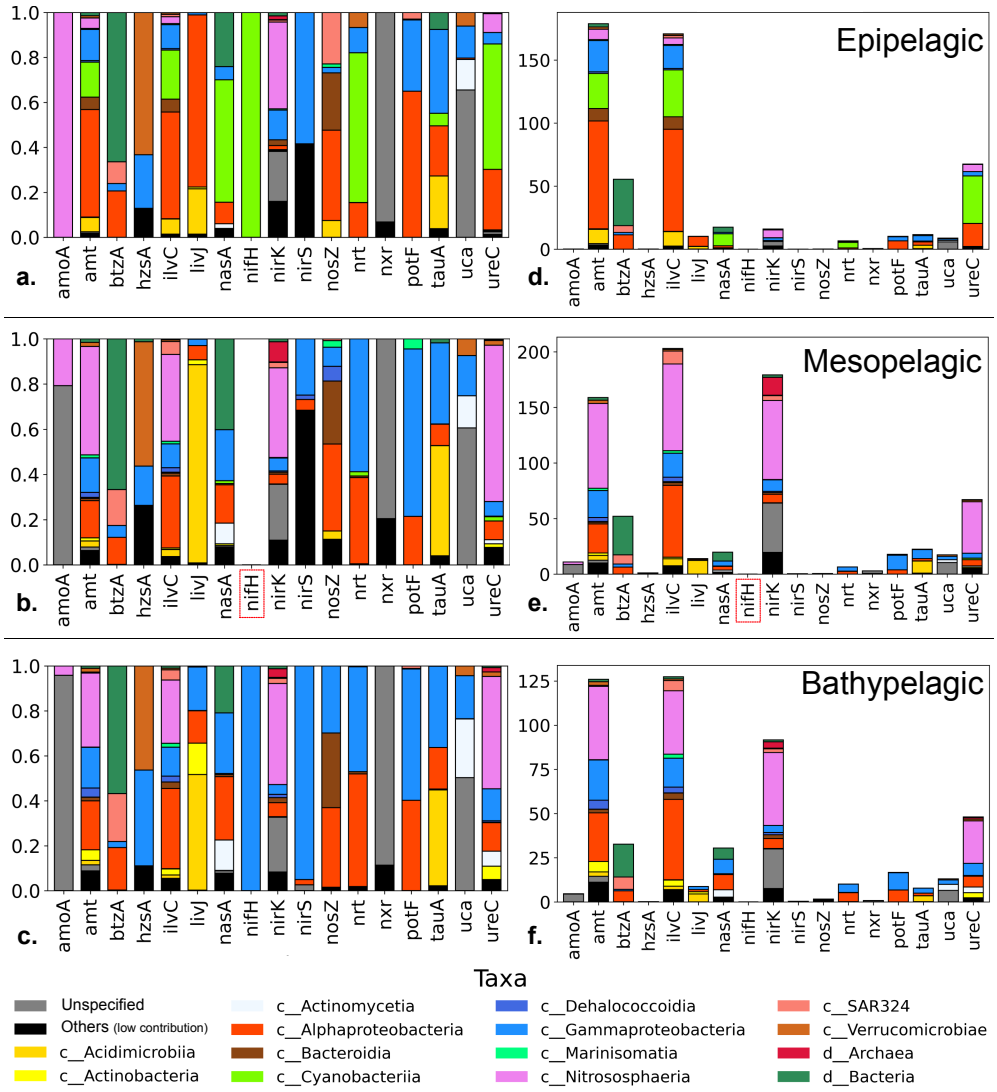

Fig. 2 Relative abundance of the studied genes in features per million (FPM) for the three water layers. In panels d-f the total counts are shown. In panels a-c the percent taxonomic composition is shown to see the contributions of the less abundant components. The most abundant genes were amt (ammonia uptake), ilvC (carboxylated aminoacid uptake, nirK (denitrification) and ureC (urea degradation). The most important changes occurred between the epi and the mesopelagic, especially in the taxonomic composition of the most abundant genes, and in the abundance of amoA and nirK genes. The nifH gene for N fixation did not appear in the mesopelagic (indicated in red) and was very rare at other depths.

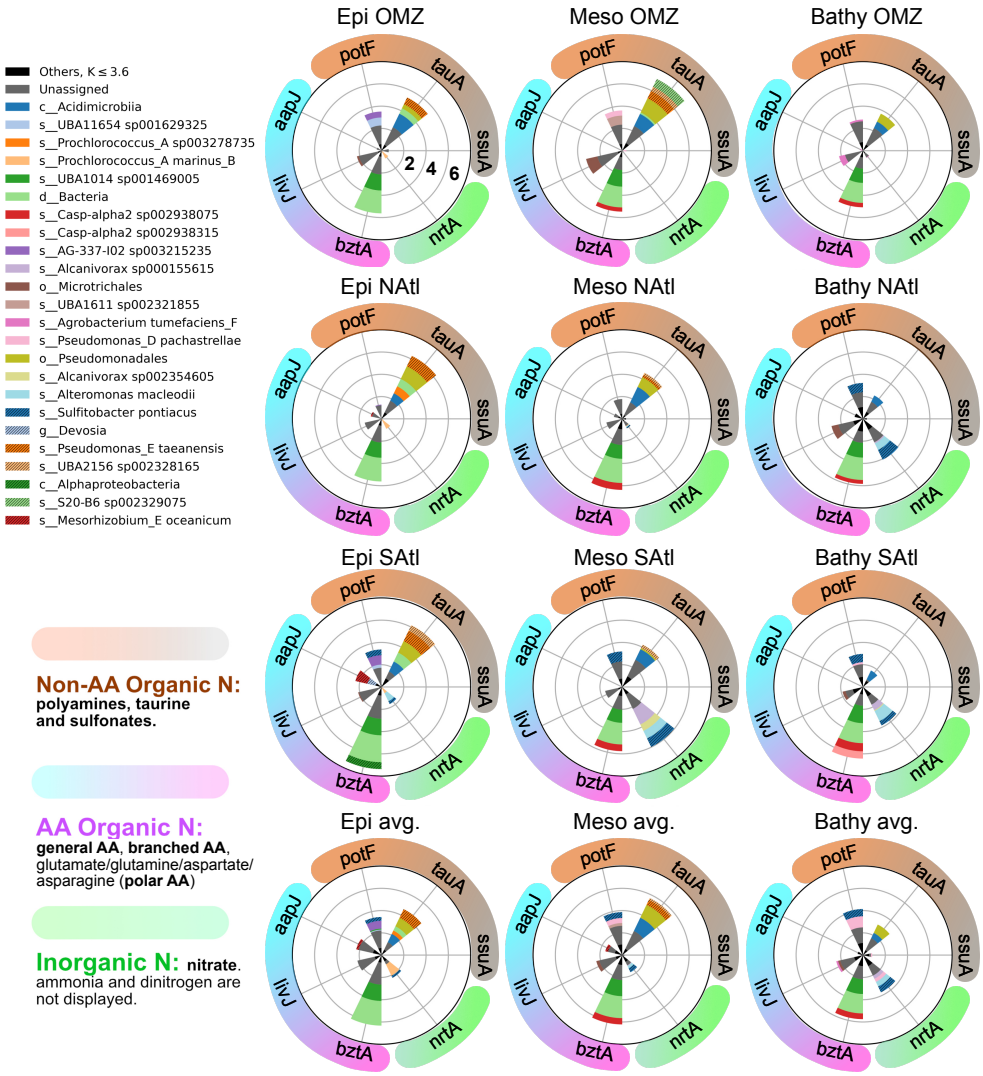

Fig. 3 Relative importance of the marker genes for N uptake guilds as represented by the log(k-values) that include components of abundance and unexpected diversity (see text). We represented the top 25 taxonomic contributors, for the sake of clarity. All other taxa were grouped under Others (black). Genes have been organized into inorganic N uptake (green color, only nirK, since amt is too abundant and nifH almost completely absent, and including them in the graphs would not allow visualization of most other genes), amino acid uptake genes (cyan to magenta), and organic non amino acid uptake genes (brown shades). The concentric circles indicate orders of magnitude. The lower row of polar graphs shows the average for all stations except the three above. This shows the general pattern of N acquisition genes was similar in most stations. Although it is not shown, ammonia uptake was always important, as amt was always the most abundant gene. The three upper rows show values for the OMZ, the Natl and the SAtl stations respectively.

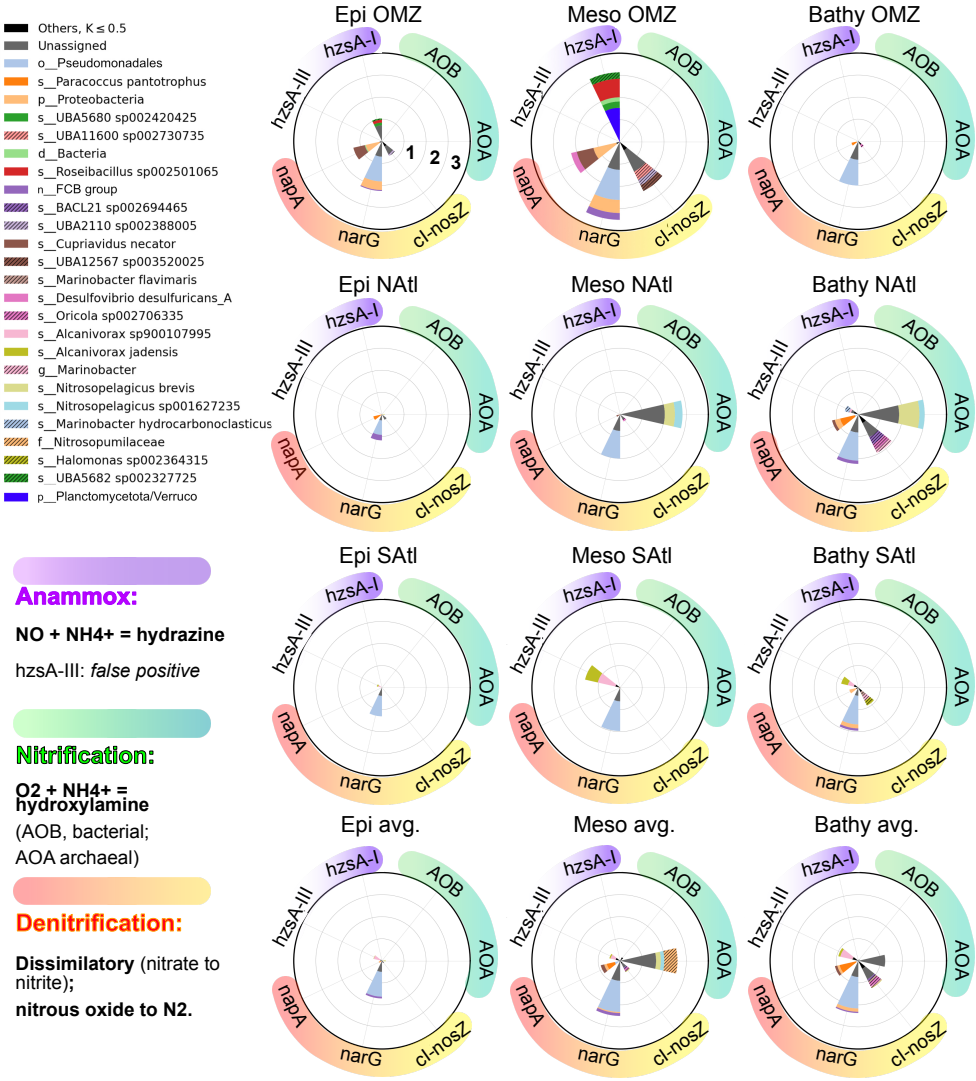

Fig. 4 Relative importance of the marker genes for N redox guilds as represented by the log(kvalues) that include components of abundance and unexpected diversity (see text). We represented the top 25 taxonomic contributors, for the sake of clarity. All other taxa were grouped under Others (black). Genes have been organized into denitrification (pink to yellow shades), nitrification (green) and putative anammox (purple shades). Differences among stations and depths were larger than for N uptake genes (see text).

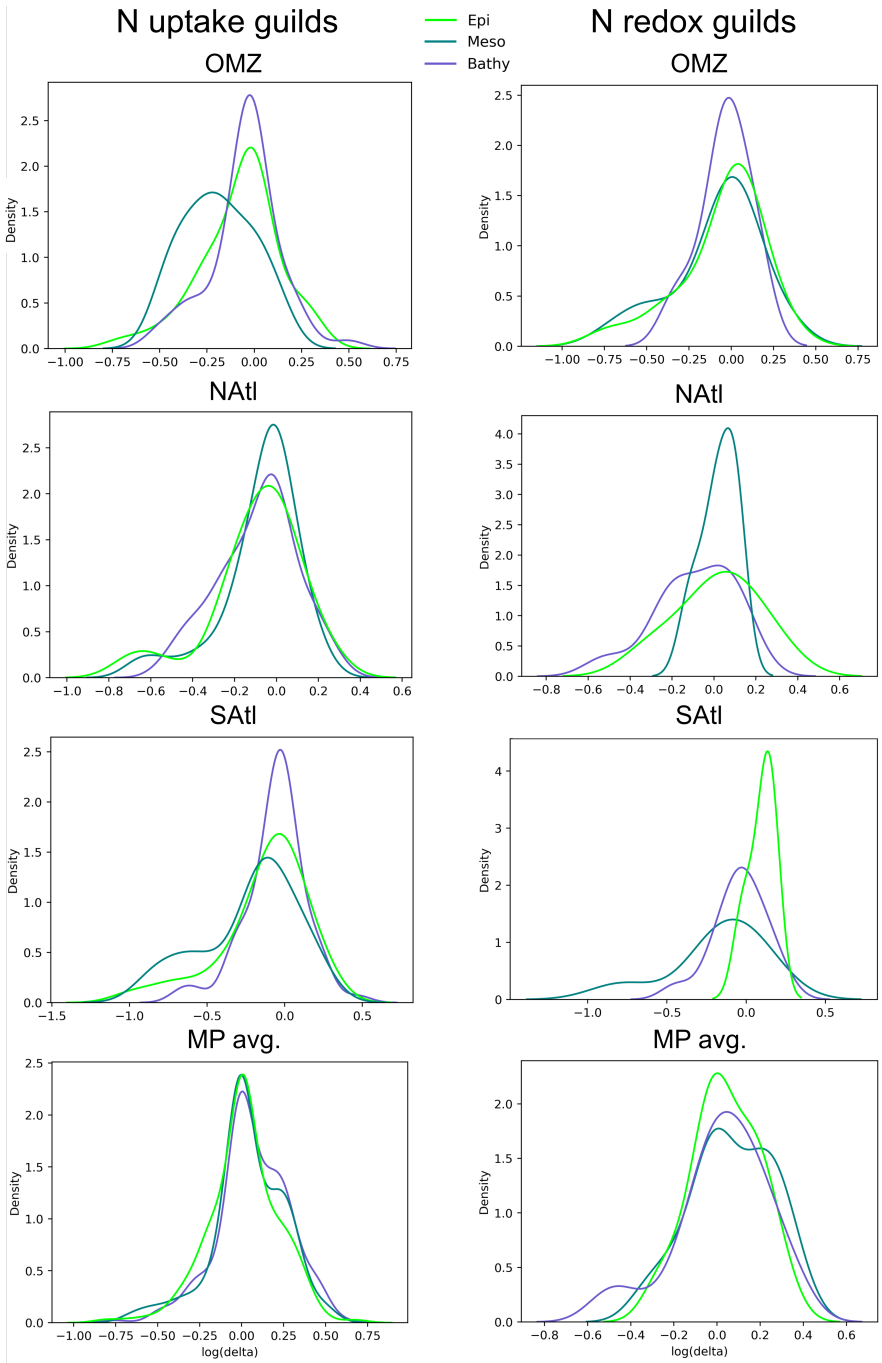

Fig. 5 Density functions for δ values for the different depths and stations analyzed. Genes and taxa with a diversity of sequences expected from their abundance will have a δ of 1 and will appear in the x axis around the zero values ($\log(1) = 0$). Those with a diversity lower than expected will appear to the left and, in contrast, those with a diversity larger than expected will appear to the right. In most cases the curves are centered around the value of zero. But in other cases there are clear departures, for example at the OMZ mesopelagic for N uptake genes.

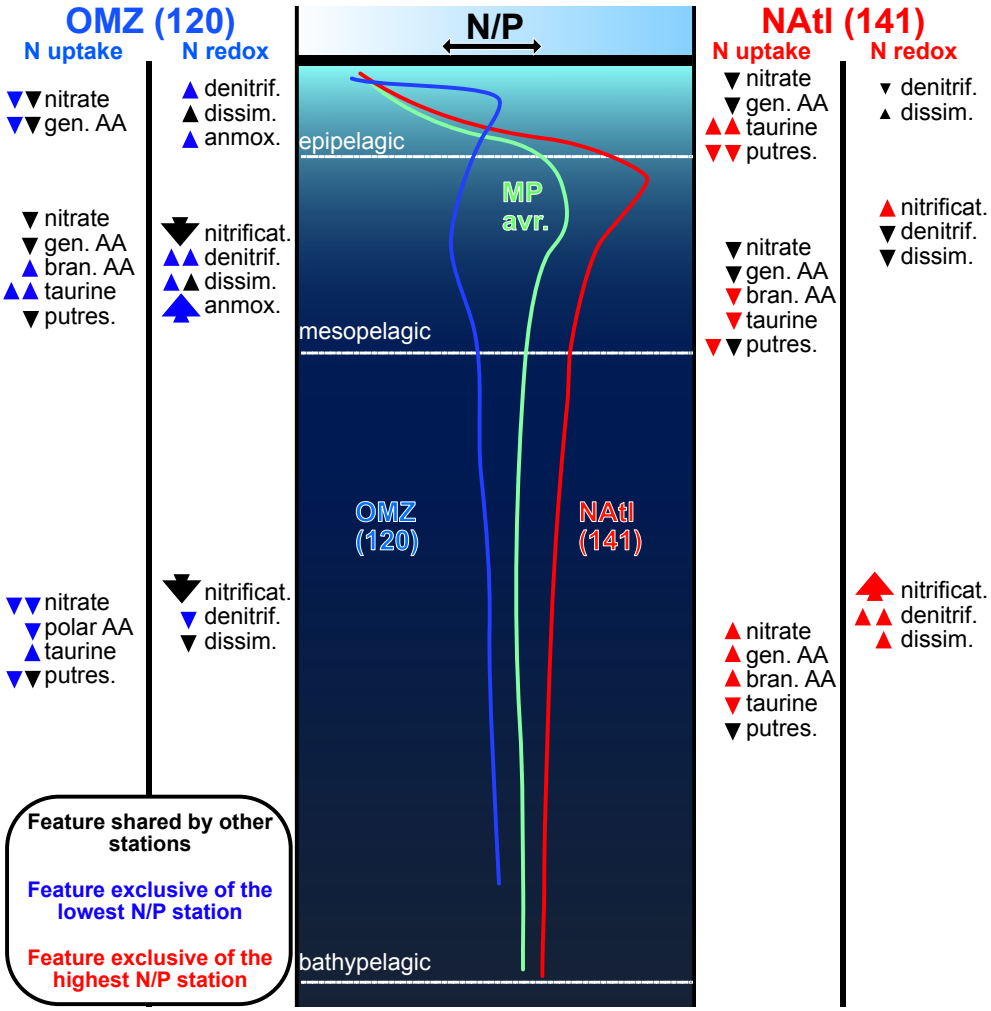

Fig. 6 Summary of the main findings. The central cartoon shows the vertical profiles of the N/P ratios for the average ocean (green) and for the stations with the lowest (station 120, OMZ in blue) and highest (station 141, Natl in red) N/P ratios. The trends for the two later are shown at the left and right of the cartoon respectively. The relative strength of the change is shown qualitatively by the number and size of the arrows. The direction of the arrows (upward or downward) indicates whether the process in question increased or decreased with respect to the average. The black arrows indicate that this change in the process was found in at least another water column. When it was exclusively found only at one of these two stations the arrows are colored. Abbreviations. Gen. AA: general L-amino acid incorporation. Bran. AA: branched amino acid incorporation. Putres.: putrescine uptake. Denitrif.: last step of denitrification + presence of the first steps of it. Dissim.: dissimilatory nitrate reduction. Anmox.: anammox.