Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities

Silvia G. Acinas, Pablo Sánchez, Guillem Salazar, Francisco M. Cornejo-Castillo, Marta Sebastián, Marta Royo-Llonch, Lucas Paoli, Shinichi Sunagawa, Pascal Hingamp, Hiroyuki Ogata, Gipsi Lima-Mendez, Simon Roux, José M. González, Jesús M. Arrieta, Intikhab S. Alam, Allan Kamau, Chris Bowler, Jeroen Raes, Stéphane Pesant, Peer Bork, Susana Agustí, Takashi Gojobori, Dolors Vaqué, Matthew B. Sullivan, Carlos Pedrós-Alió, Ramon Massana, Carlos M. Duarte & Josep M. Gasol

The deep sea, the largest ocean’s compartment, drives planetary-scale biogeochemical cycling. Yet, the functional exploration of its microbial communities lags far behind other environments. Here we analyze 58 metagenomes from tropical and subtropical deep oceans to generate the Malaspina Gene Database. Free-living or particle-attached lifestyles drive functional differences in bathypelagic prokaryotic communities, regardless of their biogeography. Ammonia and CO oxidation pathways are enriched in the free-living microbial communities and dissimilatory nitrate reduction to ammonium and H₂ oxidation pathways in the particle-attached, while the Calvin Benson-Bassham cycle is the most prevalent inorganic carbon fixation pathway in both size fractions. Reconstruction of the Malaspina Deep Metagenome-Assembled Genomes reveals unique non-cyanobacterial diazotrophic bacteria and chemolithoautotrophic prokaryotes. The widespread potential to grow both autotrophically and heterotrophically suggests that mixotrophy is an ecologically relevant trait in the deep ocean. These results expand our understanding of the functional microbial structure and metabolic capabilities of the largest Earth aquatic ecosystem.
Most of the ocean’s life is isolated from sunlight, our planet’s primary energy source. Besides living in permanent darkness, deep-ocean organisms have to cope and adapt to the high pressure and low temperature that characterize this ecosystem. This fascinating habitat represents one of the largest biomes on Earth, mostly occupied by bacteria and archaea that play a pivotal role in biogeochemical cycles on a planetary scale\textsuperscript{1,2}. Microbial metabolisms in the deep ocean have been assumed to be primarily heterotrophic, relying on organic matter exported from the sunlit layer through sinking particles (zooplankton fecal pellets, phytoplankton aggregates, and other types)\textsuperscript{3,4}. However, the high respiratory activity measured in the dark ocean is difficult to reconcile with the rates of supply of organic carbon produced in the photic layer\textsuperscript{5–7}, suggesting the existence of other sources of carbon, such as potentially non-sinking POC and in situ production by autochthonous microbial chemolithoautotrophs\textsuperscript{8}. Additional pathways that inject particles into the ocean interior (particle injection pumps; PIPs)\textsuperscript{9} can be mediated by physical processes such as subduction\textsuperscript{10,11} or biological processes, such as the daily migration of zooplankton and small fish\textsuperscript{12,13}. Chemolithoautotrophy has been considered as a possible pathway supporting the high dark ocean respiratory activity\textsuperscript{8,14–16}. Experimental rate measurements and bulk biogeochemical estimates agree with findings using single-cell genomics that ubiquitous bacterial lineages have the potential for inorganic carbon fixation\textsuperscript{17,18}, suggesting that chemolithoautotrophy may be a substantial contributor to deep-sea metabolism and may play a greater role in the global ocean carbon cycle than previously thought. Whereas inorganic carbon fixation is energetically costly\textsuperscript{19}, mixotrophy, i.e., carrying out chemolithoautotrophic inorganic carbon fixation and heterotrophy\textsuperscript{20,21}, may constitute a cost-effective strategy for microorganisms to persist in the dark ocean, although its extent in the global deep ocean is unknown.

Despite the potential importance of organic carbon derived from chemolithoautotrophy, particles likely represent the main source of reduced C to the deep ocean and constitute important hotspots of microbial activity that fuel the dark ocean food web\textsuperscript{7}. These can be fast-sinking particles, traveling through the water column in a few weeks\textsuperscript{22–24}, or buoyant or slow-sinking organic particles, which remain suspended in the deep ocean over annual time scales\textsuperscript{7}. The delivery of fast-sinking particles to the deep ocean depends on the trophic and functional structure of the surface ocean. As a result, this flux is intermittent\textsuperscript{25} and heterogeneously distributed on the spatial scale, which may lead to a heterogeneous distribution in the metabolic capacities of deep-ocean microbes across the global ocean. The diversity and biogeography of bathypelagic prokaryotic communities have recently been described at a global scale, showing that the free-living (FL) and particle-attached (PA) microbial communities differ greatly in taxonomic composition\textsuperscript{26,27} and appear to be structured by different ecological drivers\textsuperscript{28}. The lifestyle dichotomy between FL and PA prokaryotes was shown to be a phylogenetically conserved trait of deep-ocean microorganisms\textsuperscript{29}, however, the differences in the functional capacities of these two groups of microorganisms remain largely unexplored. Despite the existence of some studies at the local/regional scale\textsuperscript{29–32}, a global understanding of the ecology and metabolic processes of deep-sea bacterial and archaean microorganisms similar to that available for the upper/photic ocean is still missing\textsuperscript{33–35}. The Malaspina 2010 Circumnavigation Expedition aimed to address such knowledge gaps by surveysing bathypelagic microbes in the tropical and subtropical oceans\textsuperscript{36}. Here, we analyze 58 deep-sea microbial metagenomes from that expedition and release the Malaspina Gene Database (M-GeneDB), a valuable deep-ocean microbial community genomic dataset. In addition, we constructed the Malaspina Deep Metagenome-Assembled Genomes (MDeep-MAGs) catalog to explore the metabolic potential of the deep-sea microbiome.

Results and discussion

Gene-centric taxonomic composition of the Malaspina Gene DataBase (M-GeneDB). The 58 microbial metagenomes were sampled between 35°N and 40°S from 32 stations (St) in the North and South Pacific and Atlantic Oceans, the Indian Ocean, and the South Australian Bight (Fig. 1a) from an average water depth of 3731 m (Fig. 1b). Two different plankton size fractions were analyzed in each station representing the FL (0.2–0.8 μm) and PA (0.8–20 μm) microbial communities, the latter fraction also including bathypelagic microbial eukaryotes. A total of 195 gigabases (Gb) (6.49 × 10\textsuperscript{8} read pairs) of the data with an average of 3.36 Gb per sample were generated (Supplementary Data 1). The number of predicted genes was 4.03 million (M) from the 58 assembled bathypelagic metagenomes. M-GeneDB was first built by clustering nucleotide sequences at 95% sequence identity to remove redundancy and for consistency with the Tara Oceans gene set, yielding 1.12 M non-redundant unique sequence clusters (referred hereafter as genes). The M-GeneDB was next integrated into the recently updated Ocean Microbial-Reference Gene Catalogue of Tara Oceans (OM-RGC.v2\textsuperscript{37}) by further clustering both databases. The novelty of the M-GeneDB is represented by a total of 647,817 Malaspina-exclusive genes that account for 58% of the total genes (Fig. 1c) that were absent from the Tara Oceans global survey\textsuperscript{32}, representing a unique gene repertoire complementary to the epipelagic and mesopelagic genes reported by Tara Oceans\textsuperscript{34,37} (Fig. 1b, c). The nature of the novelty of the M-GeneDB lies in its 63% of “unknown” genes without functional annotation. The remaining 37% genes were linked to transporters, including two-component systems and ABC transporters (>97%), followed by DNA repair and recombination proteins (2.6%) and peptidases (1.8%; Supplementary Fig. 2; Supplementary Data 2). Other genes >1% were secretion systems, aminoacidic related enzymes, or quorum sensing genes (Supplementary Fig. 2). These results agree with previous findings of the prevalence of these genes in the deep ocean\textsuperscript{39–40}.

Although we acknowledge that this comparison represents a snapshot of the currently known databases, this number reflects the vertical functional stratification and dichotomy between photic and aphotic microorganisms, as well as the differences between the mesopelagic and the bathypelagic. Despite the role of sinking particles in delivering epipelagic microbes to the deep-sea\textsuperscript{41}, the presence of endemic bathypelagic microorganisms has been described in several of the Malaspina sampled stations\textsuperscript{41}. On average 61 ± 14% SD of the predicted genes in each sample were exclusively found in the present dataset, which highlights the unique gene content of the bathypelagic microbiome (Fig. 1c and Supplementary Data 1). Each sample contained 14 ± 9% specific genes not found in any other Malaspina samples (Fig. 1c and Supplementary Data 1). Station St62 in the Indian Ocean, sampled at 2400 m, showed the highest fraction of sample-specific genes with 43% of the total being also highly different in terms of taxonomic community composition\textsuperscript{26} (Supplementary Fig. 3). This sample, together with other four stations located in Brazil (St32), North Atlantic American (St134), and Guatemala basins (St121), that harbored more than 30% of sample-specific genes, were all from the PA size fraction and associated with circumpolar deep water and North Atlantic Deep Water masses (Supplementary Data 1).

The taxonomic affiliation of the genes in the M-GeneDB indicated that most of them belonged to the bacteria and archaea
domains, with a minor representation of viruses in both size fractions and eukaryotes that were mostly found in the PA fraction (Supplementary Fig. 1). Nevertheless, a significant proportion (~25–30%) of the genes in both size fractions could not be assigned to any domain (Supplementary Fig. 1). Microbial taxonomy (i.e., small eukaryotes (Supplementary Fig. 3a)), prokaryotes (Supplementary Fig. 3b), giruses (Supplementary Fig. 3c), and viruses (Supplementary Fig. 3d) in the bathypelagic samples was evaluated using different marker genes extracted from the metagenomes (Supplementary Data 3, Supplementary Data 6, see Supplementary Discussion). The identified diversity concurred with previous results based on 18S42 and 16S rRNA26 PCR amplicons from the same samples. Identification of nucleocytoplasmic large DNA viruses (NCLDV)s revealed their ubiquity in both size fractions and in all ocean basins (Supplementary Fig. 3c). The dominant NCLDV in the deep ocean were Megaviridae (76% and 63% in the FL and PA fractions, respectively). This result contrasts with the lower proportion (36%) of Megavirida in the sunlit ocean43.

**Functional architecture of the deep-ocean microbiome.** We evaluated the effect of the different oceanic regions, basins, and lifestyles (FL or PA) on determining the bathypelagic prokaryotic functional structure (Fig. 2). Our results identified two main functional groups of samples corresponding to PA and FL communities (Fig. 2a). This pattern was coherent regardless of the functional classifications used: the Kyoto Encyclopedia of Genes and Genomes Orthologs (KO; Fig. 2), clusters of orthologous groups (COG), protein families (Pfam), or Enzyme Commission numbers (EC; Supplementary Fig. 4). The PA and FL lifestyle explained 20.6% of the variance followed by ocean basins (16.6%) and oceans (7.4%; PERMANOVA test, P < 0.001). Although previous studies have also shown contrasting gene repertoires for FL and PA microbial communities, they were limited to a few studies at the local scale in the photic ocean47,48 or to the oxygen minimum zone (OMZ) off the coast of Chile32 and a global comparison had not been presented.

These findings highlight that the main factor that functionally structures the microbial communities in the bathypelagic deep ocean is community lifestyle rather than their geographic origin, although differences among the different oceanic basins (Fig. 2b) were also observed. Thus, FL and PA prokaryotic microbial communities of the deep ocean not only represent different taxonomic groups as previously recognized26,27 but also correspond to lifestyles with consistently different functional traits.

To explore the potential metabolic differences between FL and PA prokaryotic microbial communities (Figs. 3 and 4), a selection of marker genes (Supplementary Data 7) related to...
potential relevant pathways in the deep ocean related to inorganic carbon fixation, nitrogen, sulfur, methane, and hydrogen metabolisms were searched in the bathypelagic metagenomic dataset (Fig. 3). Out of these key marker genes, 49 KOs were present in the M-GeneDB (Supplementary Data 8). Among them, the key genes of four different inorganic CO₂-fixation pathways.

The Calvin–Benson–Bassham (CBB) cycle, identified by the ribulose-bisphosphate carboxylase large subunit (RuBisCO, K01601: rbcl) was widely distributed in the dataset followed by the 3-HP (K14468: mcr, K08691: mcr), whereas the archaeal 3-hydroxypropionate–4-hydroxybutyrate cycle (K15039), and the rTCA cycle (K15230: aclA, K15231: aclB) displayed a narrow
distribution restricted to a single station each: St62 and St53, respectively, both in the Indian Ocean (Fig. 3). RuBisCO occurred in 48% of the samples (Supplementary Data 8), and on 1.3% of the potential cells on average (based on recA normalization) with two significant deviations observed in the PA fractions at St67 (South Australian Bight, Indian Ocean) and St134 (North American basin, Atlantic Ocean) where it peaked up to 12% of the cells (Fig. 3). However, there were no significant differences between both size fractions (Wilcoxon test, P value = 0.135).

Nitrification through ammonia\textsuperscript{15,50,51} and nitrite oxidation\textsuperscript{18} have been postulated as the main sources of energy for carbon fixation in the dark ocean. For ammonia oxidation, we looked for

\textbf{Fig. 4 Comparison of metabolic pathways in the free-living and particle-attached microbial communities from the bathypelagic ocean.} Normalized gene abundance (reads/recA reads) of 49 marker genes indicative of different autotrophic carbon-fixation pathways, nitrogen, sulfur, methane, and hydrogen metabolisms in the Malaspina Gene DataBase. Gene abundances from KOs belonging to the same pathway and KEGG module level have been plotted together (Supplementary Data 7). From top to the bottom: Reductive citrate cycle (rTCA): K15230, K15231; hydroxypropionate-hydroxybutyrate cycle: K15039; dicarboxylate-hydroxybutyrate cycle: K15016; Calvin cycle: K00855, K01601 and K01602; 3-hydroxypropionate bi-cycle: K14468, K14470 and K09709; CO oxidation (C1 metabolism): K02520, K02519 and K02518; H2-oxidation: K00436; methane oxidation: K16158; coenzyme M biosynthesis: K05979, K06034, K08097 and K13039; C1 metabolism/methanogenesis: K00320, K00200, K00201, K00202, K00672, K03390, K14083, and K00577; nitrification/methane oxidation: K02588; nitrification/denitrification/dissimilatory nitrate reduction: K00370 and K00371; dissimilatory nitrate reduction (DNRA): K0262 and K0263; denitrification/dissimilatory nitrate reduction: K00374, K02567, and K02568; denitrification: K00368, K15864, K04561, K02305, and K00376; assimilatory nitrate reduction: K00367 and K00372; dissimilatory sulfate reduction: K00394, K00395 and K11181. Wilcoxon tests were done to test for significant differences between the particle-attached (PA) and free-living (FL) assemblages and significant (P value < 0.05) differences are labeled with asterisks. FL (empty boxes) and PA (filled boxes) bathypelagic microbial communities are shown next to each other.
the KEGG ortholog K10944 (pmoA-amoA) corresponding to the marker gene methane/ammonia monooxygenase subunit and the protein family PF12942 to identify the archaeal ammonia monoxygenase subunit A. The abundances of both markers across samples were positively correlated (Spearman correlation, \( r = 0.69, P = 2.55\times10^{-9} \)), and they were found in \( \sim 36\% \) of our samples and in \( 6\% \) of microbial cells, mostly in FL microorganisms (Wilcoxon test, \( P < 0.005 \) (Supplementary Fig. 5). For nitrite oxidation, the nitrate reductase/nitrite oxidoreductase (K00370/K00371) was found in \( 81\% \) of the samples and in \( \sim 4\% \) of the microbial cells, although this enzyme could also participate in other metabolic processes as in dissimilatory nitrate reduction and denitrification (Fig. 3). Thus, exploring the co-occurrence of other key genes of each pathway in metagenomic bins is necessary for their validation as shown below. Other relevant nitrification genes, such as the hydroxylamine dehydrogenase (K10535: hao) or the key enzyme for the anaerobic ammonia oxidizers (K20932/K20935: hdn; hydrazine dehydrogenase; ammomox bacteria), normally present within the oxygen minimum zones in the mesopelagic ocean, were absent from our bathypelagic metagenomic dataset. This reflects different biogeochemical processes occurring in the anoxic mesopelagic OMZ and the oxic bathypelagic oceans.

\( \text{H}_2 \) and CO oxidation were explored as potential alternative energy sources\(^5\). \( \text{H}_2 \)-oxidation has been described in hydrothermal vents\(^5\) or subsurface microbial communities\(^5\) but we also found it in \( 24\% \) of the samples (Fig. 3), mostly in PA microorganisms (Wilcoxon test, \( P < 0.005 \); Fig. 4). These results expand the ecological niches of microbial \( \text{H}_2 \) oxidizers in the bathypelagic ocean, probably associated with particles providing anoxic microenvironments where \( \text{H}_2 \) production by fermentation is favored. The oxidation of \( \text{CO} \) is catalyzed by CO dehydrogenase (CODH; cox genes)\(^5\) and has been associated with Actinobacterota, Proteobacteria, and members of Bacteroidota and Chloroexota phyla.\(^5\) The ubiquitous distribution of \( \text{CO} \) oxidation by cox genes (K03518: cox8; K03519: cosM and K03520: coxL) was notable since it was detected in \( 87\% \) of the samples and with a high abundance (average \( 20\% \) of the microbial cells; Fig. 3), mostly in FL prokaryotes (Wilcoxon test, \( P < 0.005 \) (Fig. 4), pointing to \( \text{CO} \) oxidation as an important energy supplement for heterotrophs in the deep ocean.

Dissimilatory nitrate reduction to ammonium (DNRA) was also detected. Due to the absence of the periplasmic pentahem cytochrome \( c \) nitrite reductase \( \text{nrfA} \)\(^5\) in our dataset (Supplementary Data 7), nitrate reduction to ammonium seemed to be catalyzed by the cytoplasmic NADH-dependent nitrite reductase \( \text{nirB} \) or by its two-subunit variant \( \text{nirBD} \) (K00362: \( \text{nirB} \) -K00363: \( \text{nirD} \))\(^5\). The potential DNRA and other metabolisms such as denitrification (K00368: \( \text{nirK} \); K02305: \( \text{nirG} \); K04561: \( \text{nirB} \); K00376: \( \text{nirZ} \)) and sulfate reduction (K00394: \( \text{aprA} \) -K00395: \( \text{aprB} \)) were also present. Up to \( 27\% \) of the samples contained denitrification-related genes and \( 90\% \) of the samples displayed marker genes of sulfate reduction (Fig. 3 and Supplementary Data 8). The prevalence of such metabolisms in well-oxygenated environments inside organic aggregates or particles, where intense respiration may result in local \( \text{O}_2 \) exhaustion\(^6\). This is supported by the finding that both the assimilatory and dissimilatory nitrate reduction pathways were enriched in the PA fraction (Wilcoxon test, \( P < 0.005 \); Fig. 4). While the potential for DNRA was present in most of the samples at abundances that reached up to \( 34\% \) of the potential microbial cells (Fig. 3), denitrification was less abundant (between \( 0.7 \) and \( 8\% \) of microbial cells) despite being widely distributed. Finally, dissimilatory sulfate reduction genes were found in most of the FL and PA microbial communities in \( 5\% -13\% \) of the cells (Fig. 3). Overall, we found that the CBB cycle was distributed in both size fractions, whereas the \( \text{CO} \) oxidation or ammonia oxidation were enriched in FL and \( \text{H}_2 \) oxidation, and conversely DNRA, were enriched in PA microbial communities. Our results not only highlight the distribution of the main potential biogeochemical processes in the bathypelagic ocean, some of them presenting a patchy distribution, but also attribute specific metabolic pathways to either the deep-ocean FL or PA prokaryotes.

### Diversity and novelty of the Malaspina Deep MAGs catalog

Co-assembly of the 58 bathypelagic metagenomes allowed the reconstruction of 619 non-redundant bins with \( 57.2\% \) mean genome completeness, accounting for a total of \( 1.4 \) Gbp. A total of 317 of these bins had \( \geq 50\% \) genome completeness values and \( < 10\% \) contamination and fulfilled the quality standards\(^6\) to be considered as medium or high-quality Metagenome-Assembled Genomes (MAGs). These 317 MAGs, with \( 84.2\% \) average genome completeness, represented a total of \( 936 \) Mbp. This dataset is referred to here as the Malaspina Deep MAGs catalog (MDeep-MAGs). A total of \( 298 \) bacterial MAGs were taxonomically assigned to 22 phyla and 19 archaeal MAGs were assigned to 4 phyla based on the GTDB taxonomy (Fig. 5). In addition, two low-quality bins (0046 and 0224) were assigned to euarcheota, potentially to fungal taxa. Proteobacteria (\( n = 149 \)) followed by Bacteroidota (\( n = 30 \)) and Chloroflexota (\( n = 19 \)) were the most abundant bacterial phyla in the MDeep-MAGs collection, while for archaeal MAGs Thermoplasmatota (\( n = 9 \)) and Crenarchaeota (\( n = 6 \)) had the most representatives (Fig. 5a). The more abundant bacterial classes were Gammaproteobacteria (\( n = 82 \)) and Alphaproteobacteria (\( n = 67 \)). The MDeep-MAGs included a remarkable taxonomic novelty with \( \geq 68\% \) and \( 58\% \) of novel species within the archaea and bacteria MAGs, respectively. Within bacteria, MAG0213 was assigned to a novel Class of the Latecibacterota phylum, one MAG may represent a novel order and five MAGs could represent distinct novel families (Fig. 5b and Supplementary Data 9). In the case of archaea, we found MAG0485 as potentially representing a novel family within the Nanoarchaeota phylum (Fig. 5b).

Interestingly, the MDeep-MAGs recruited \( 32\% \) of the reads of the total free-living fraction bathypelagic metagenomes and \( \sim 20\% \) of the reads in the PA fraction (Fig. 5c). These differences may suggest that a larger proportion of diversity is missing from the PA fraction. One possibility is due to the presence of picoeukaryotes since protists, which have larger genomes, are more fragmented in the metagenome and are more difficult to bin into MAGs, as suggested by the lower individual assembly sizes of PA compared to FL (\( 22.4 \pm 13.6 \) and \( 36.9 \pm 17.1 \) Mbp, respectively; Supplementary Data 1) obtained with the same sequencing depth. This lower read mapping coming from these large-size fraction samples may be due to the lower genome reconstruction of picoeukaryotes. Also, it has been shown that the prokaryotic phylogenetic diversity per OTU and the mean nearest taxon distance (MNTD) is higher for the PA fraction\(^2\) and this may affect the lower mapping rate on the genomes present in this fraction.

Overall, these recruitments were higher than those reported for the photic layer of the global ocean Tara Oceans dataset (6.84\% of the metagenomic reads)\(^6\) and specifically for the Mediterranean Sea, which showed average mapping rates of 14\% of the metagenomic reads for different microbial size fractions\(^6\). Such discrepancy may be due to probable differences in sequencing depth and methodological variations in the co-assembly, filtering, and mapping between studies.

MAG completion estimates based on domain-specific single-copy core genes\(^8\) ranged from 50.3 to 100\% (Supplementary
Genome-resolved metabolic capabilities of the bathypelagic ocean microbiome. The MDeep-MAGs catalog represents the most extensive genome dataset from the bathypelagic ocean built to investigate the distribution and functional capability of deep-ocean microorganisms (Fig. 6). In the following section, we describe the most important features in the dataset.

Non-cyanobacterial diazotrophs (NCDs). The ecological relevance of non-cyanobacterial diazotrophs (NCDs) in the deep ocean remains unclear as the availability of carbon substrates seems unlikely to support the costly energetic demands of N₂ fixation. However, NCDs are widely distributed across oxygenated oceans and are phylogenetically diverse. New genotypes associated with Planctomycetota and Proteobacteria were recently found in the surface global ocean, and some of them were actively transcribed even at mesopelagic depths. The diversity of NCDs in the deep ocean is mostly known from the detection of the nifH gene. Three NCD MAGs were reconstructed in the Malaspina MDeep-MAGs dataset that harbored the nifH gene and other structural genes of the nif operon such as nifK, nifD, and other structural genes that were almost complete genomes (>94% completeness) that were detected exclusively in the PA fraction. The third one was a Gamma-proteobacteria (MAG0081) in the genus *Ketobacter* present in both size fractions (Fig. 7). The presence of the Alpha-proteobacteria diazotrophic MAGs in the PA fraction fits well with the finding of diazotrophic bacteria in sinking mesopelagic particles at the North Pacific Subtropical Gyre. Nevertheless, the distribution of these three MAGs was restricted to a limited number of samples (between 5 and 9% of the total).

MAG0509 is closely related to *Salipiger thiooxidans*, a sulfur-oxidizing lithoheterotrophic bacterium isolated from the Black Sea. Interestingly, in addition to the nifH and other structural genes from the nif operon (nifK), this MAG has two genes of the RuBisCo (forms I and IV), the cox genes for CO oxidation and a soxY gene for thiosulfate oxidation, pointing to potential chemolithoautotrophic diazotrophy.
These results may reflect a higher metabolic versatility within this taxon, with the presence of previously undetected chemolithoautotrophic diazotrophs in the bathypelagic ocean. The other Alphaproteobacteria genome related to Novosphingobium (MAG0177) was also intriguing: although members of this genus are known to be metabolically versatile and usually associated with the biodegradation of aromatic compounds, they have been commonly isolated from sites impacted by anthropogenic activities. The only known species capable of N₂ fixation is Novosphingobium nitrogenifex, isolated from a pulp and paper wastewater bioreactor with the ability to accumulate polyhydroxyalkanoate. MAG0177 displayed a wide range of metabolic versatility within this taxon, with the presence of previously undetected chemolithoautotrophs in the bathypelagic ocean. The other taxon, with the presence of previously undetected chemolithotrophic diazotrophs in the bathypelagic ocean. The other taxon, with the presence of previously undetected chemolithoau-

**Fig. 6 Metabolic potential in selected metagenome-assembled genomes (MAGs) from the bathypelagic ocean.** A total of 25 MDeep-MAGs were selected based on the presence of metabolic pathways with potential for chemolithoautotrophy, mixotrophy, and nitrogen fixation (non-cyanobacterial diazotrophs, NCDs). AOA ammonia-oxidizing archaea, SOB sulfur-oxidizing bacteria, NCDs non-cyanobacterial diazotrophs, NOB nitrite-oxidizing bacteria, AOB ammonia-oxidizing bacteria. Their taxonomic assignment at the maximal possible resolution is shown at the top, followed by the occurrence of each MAG in samples of the Malaspina Gene DataBase, and the genome completeness (%) of each MAG. On the right, metabolic pathways involved in inorganic carbon fixation (green), sulfur (red), nitrogen (blue) are shown as well as the KOs that participate in these pathways. At the bottom, the percentage module completeness for each pathway is coded by color intensity.

These results may reflect a higher metabolic versatility within this taxon, with the presence of previously undetected chemolithoautotrophic diazotrophs in the bathypelagic ocean. The other Alphaproteobacteria genome related to Novosphingobium (MAG0177) was also intriguing: although members of this genus are known to be metabolically versatile and usually associated with the biodegradation of aromatic compounds, they have been commonly isolated from sites impacted by anthropogenic activities. The only known species capable of N₂ fixation is Novosphingobium nitrogenifex, isolated from a pulp and paper wastewater bioreactor with the ability to accumulate polyhydroxyalkanoate. MAG0177 displayed a wide range of xenobiotic biodegradation pathways, such as those for xylene, toluene, and benzoate degradation (Supplementary Data 10).

Finally, the gammaproteobacterial NCD MAG was related to genus Ketobacter (MAG0081), with Ketobacter alkanivorans as the only species isolated from seawater impacted by an oil-spill accident, and with capacity for degradation of alkanes. We observed that alkane hydroxylase (alkB) and haloalkane dehalogenases (dhaA) genes were also detected in this genome (Supplementary Data 10) and therefore it may likely be capable of degrading synthetic haloalkanes, some of them used to make Nylon filament, fiber, and plastics and other recalcitrant compounds.

The phylogenetic analyses of these three *nifH* gene variants from our NCDs MAGs confirmed their relationship with Alphaproteobacteria and Gammaproteobacteria (Supplementary Fig. 6). MAG0081 clustered with the phylogenetic group I described by Delmont et al. of Gammaproteobacteria and matched at 100% identity with other MAGs from the Tara Oceans expedition. The other MAGs were related to the Alphaproteobacteria: the MAG0177 grouped with Novosphingobium sp. BW1 while the MAG0509 was related to Yangia, a genus of the Rhodobacteraceae.

The discovery of novel NCDs genomes from the bathypelagic ocean with presence in the PA fraction reinforces the idea of nutrient-rich sinking particles from the photic ocean as potential...
Chemolithoautotrophy and mixotrophy. The ecological relevance of chemolithoautotrophy in the deep ocean is well known from contrasting oceanic regions, such as the North Atlantic mesopelagic and bathypelagic,8,81, the central Mediterranean Sea82,83, or the Arctic Ocean21,51. Bacteria and archaea in the oxygenated water column of the dark ocean use a variety of reduced inorganic compounds, such as hydrogen, thiosulfate/sulfide, and ammonia, as energy sources8,17,21,53. However, it is unclear which are the key microbes involved and what is their prevalence in the global bathypelagic ocean. Therefore, we investigated the metabolisms associated with ammonia, sulfur, and nitrite oxidation that have been reported to be relevant in the bathypelagic ocean17,21,51 and other less explored metabolisms such as H₂-oxidation and CO oxidation. We found two ammonia-oxidizing archaea (AOA), three potential sulfur-oxidizing bacteria (SOB), one ammonia-oxidizing bacteria (AOB), and one nitrite-oxidizing bacterium (NOB) (Fig. 6).

The archael AOA (MAG0093, MAG0605) belonged to different species of *Nitrospumilus*, found only in the free-living fraction (Fig. 7) and with an occurrence of 12% of the samples (Fig. 6 and Supplementary Data 10). Phylogenetic analyses of the *amaA* genes in the AOA MAGs showed that they were closely related to other deep-sea *amaA* sequences (Supplementary Fig. 5). In both AOA MAGs, we detected the key enzyme for the archaeal 3-hydroxypropionate–4-hydroxybutyrate autotrophic carbon dioxide assimilation pathway (K14534: *abfD*), found only in the free-living fraction. The two bacterial SOB (MAG0179, MAG0595) taxonomically associated with genus Sulfurimonas, displayed a narrow distribution at a single station (St53 in the Madagascar basin) and were enriched
in the free-living fraction while the distribution of the SOB related to Gammaproteobacteria GCA-2726415 (MAG00570) was limited to the Guatemala and Pacific basins in both size fractions (Fig. 7). However, we did not find in these SOB MAGs any key gene for inorganic carbon fixation, so their chemolithoautotrophic potential remains unknown. The potential AOB (MAG0564) related to MarineAlpha9-Bin7 harbored the key genes for nitrification (hao) and for CO oxidation (coxL; Fig. 6) and displayed a wider distribution (48% of the samples) across the Atlantic, Pacific, and Indian Oceans, mostly restricted to the free-living fraction (Fig. 7 and Supplementary Data 10). Finally, the bacterial NOB (MAG0176), related to the family Nitrospinaeae, was present in both size fractions in nearly all samples (Fig. 7).

The prevalence of the potential for the oxidation of CO by CO dehydrogenase (CODH; rbcL genes) in the bathypelagic ocean fits with the findings of a recent survey in which the rbcL gene was widely distributed among aerobic bacteria and archaea in many terrestrial and marine environments. We found a total of 90 (28%) MDeep-MAGs containing the rbcL (K03520) gene, 46% of which also had the RibBiso gene (large-chain rbcL) of the CCB cycle, supporting their potential for autotrophy. However, only two MAGs contained representative genes of H2-oxidation (K00436: hoxH gene; Fig. 6): one Alphaproteobacteria of the genus Ahrensia (MAG0166) and one Gammaproteobacteria of the Kotebacter genus (MAG0081) that displayed different biogeography (Fig. 7). The Ahrensia MAG also contained the RibBiso gene for inorganic carbon fixation and it was present in over 40% of the samples (Fig. 6 and Supplementary Data 10) with a peak in the North American basin (ST131 and ST134), being present in both size fractions, although more abundant in the PA microbial communities (Fig. 7).

The presence of some of the inorganic carbon-fixation pathways, such as the CCB and the tRCA cycle associated with nitrite-oxidizing bacteria has been reported for the dark oceans, particularly for the mesopelagic and to a lesser extent in the bathypelagic ocean, by several studies combining single-cell genomics with fluorescence in situ hybridization and/or microautoradiography. Yet, the prevalence of the different inorganic carbon-fixation pathways in the bathypelagic ocean at a broad geographical scale and at genome level was unknown.

The marker gene RibBiso of the CCB cycle (K01601: large-chain rbcL) was found in 18 metagenomic bins, of which 15 are considered MAGs (totaling 4.7% out of the 317 MDeep-MAGs), whereas only MAG0084, an Alphaproteobacteria and MAG0176 belonging to Nitrospinae outputted marker genes related to alternative inorganic carbon-fixation pathways, such as the 3-HP and the tRCA cycle, respectively (Fig. 6). Two archaeal genomes (MAG0093 and MAG0605) related to Nitrosopumilus had the key gene of the hydroxypyropionate–hydroxybutyrate cycle pathway for inorganic carbon fixation (K14534: abfd; 4-hydroxybutyryl-CoA dehydratase/vinylacetyl-CoA-delta-isomerase; Supplementary Data 10).

The presence of representative genes of the 3-HP such as the key enzyme malyl-CoA lyase (K08691: mcd; 84), the 2-methylfumaryl-CoA isomerase (K14470: mcd) and 3-methylfumaryl-CoA hydratase (K09709: mcd) together with the bican carbonate carboxylase enzymes (K01961: accC, acetyl-CoA carboxylase) was surprising since this pathway was assumed to be present only in Chloroflexaceae. MAG0084, taxonomically related to Thalassospiracoccus halodurans with 98% genome completeness, displayed 67% of the 3-hydroxypropionate pathway KOs (13 out of 20), although their role in autotrophy remains uncertain. This genome was present in ~10% of the samples (Fig. 6, Supplementary Data 10) in both the FL and PA fractions (Fig. 7). Individual genes of the 3-HP pathway have recently been found in the uncultured deep-ocean SAR202 clade genomes related to Chloroflexi, although the role of the 3-hydroxypropionate bi-cycle in this clade has been linked to the assimilation of intermediate metabolites produced by the degradation of recalcitrant dissolved organic matter rather than CO2 fixation.
either FL or PA prokaryotes, expand our view of the metabolic seascape of the deep-ocean microbiome.

**Conclusions.** The global metagenomic assessment of the deep-ocean microbiome with the creation of the Malaspina Gene DataBase and Malaspina Deep MAGs catalog uncovers potentially novel orders and classes of deep-ocean microorganisms and describes potentially relevant biogeochemical processes of the deep-ocean microbiome in the tropical and subtropical bathypelagic oceans. Our results show metabolic differentiation reflected in contrasting functional gene repertoires, between the FL and the PA prokaryotic assemblages with also a certain degree of functional patchiness across oceans and basins. Our study also provides evidence for diverse metabolic strategies in the deep ocean. The widespread distribution of different autotrophic pathways in the deep ocean and the prevalence of alternative energy sources such as CO oxidation, sulfur oxidation, and H₂ oxidation, supports the role of a multitude of autotrophic processes in subsidizing the heterotrophic metabolism supported by export flux from the photic layer. These autotrophic processes depend on inorganic compounds photosynthetically reduced in the upper ocean and therefore do not constitute primary production in a strict sense. Yet, these autotrophic processes channel additional energetic resources from the upper ocean into the deep-oceanic microorganisms, allowing respiratory demands in the deep sea over those supported only by particulate organic fluxes from the photic layer.

**Methods**

**Sample collection and DNA extraction.** A total of 58 water samples were taken during the Malaspina 2010 expedition (http://www.expedicionmalaspina.es) corresponding to 32 different sampling stations globally distributed across the world’s tropical and subtropical oceans (Fig. 1a). We focused on the samples collected at the depth of 4000 m, although a few samples were taken at shallower depths, all within the bathypelagic realm (average depth: 3731 m ± 495; standard deviation). Two different size fractions were analyzed in each station representing the FL within the bathypelagic realm (average depth: 3731 m ± 495; standard deviation). Additional metadata were collected during the expedition including environmental variables (salinity, potential temperature, and oxygen concentration), sampling station coordinates (latitude, longitude, and depth), and auxiliary data for every sample (filter size, ocean basin, and water mass; Supplementary Data 1).

**Statistics and reproducibility.** For every functional abundance table, a subsampled equivalent table was constructed in order to avoid biases due to the varying sequencing depth between samples. Subsampling was performed by generating a randomly rarefied table without replacement from the original one with the "rrate" function in vegan45 package within R software46.

**Generation of the Malaspina Gene Database (M-GeneDB).** All 3,872,410 predicted coding sequences larger than 100 bp from each assembled metagenome were pooled and clustered at 95% sequence similarity and 90% sequence overlap in the smaller sequence using cd-hit-est47 v.4.6 using the following options: -c 0.95 -T 0 -s 0.9 -g 1 -r 1 -d 0 to obtain 1,155,269 non-redundant gene clusters (from now on referred simply as genes). These gene clusters were aligned to UniRef100 (release 2019-10-16) with diamond blasts49 (v.0.9.22; e-value 0.0001). The least common ancestor taxonomic assignment of UniRef100 best matches was obtained from NCBI’s taxonomy database50 (release 2020-01-30). In order to explore the novelty of the M-GeneDB, we clustered it with the 46,775,154 non-redundant sequences from the Tara Oceans Microbial Reference Gene Catalog version 2 (OM-RGC.v2)52 using cd-hit-est-2d43 v.4.6 with the following options: -c 0.95 -T 48 -M 256000 -G 0 -aS 0.9 -g 1 -r 1 -d 0 to obtain a final catalog of 47,422,971 genes.

**Taxonomy of protist, prokaryotes, and viruses from metagenomic reads**

Protist 18S miTags were extracted from the metagenomes and subsequently analyzed following Logares et al.93. These miTags were mapped at 97% similarity using Uclust101 to the PR2 database102 that was pre-clustered at 97% similarity (Supplementary Data 2).

**Bacteria and archaea.** Prokaryotic taxonomical tables were downloaded from IMG [Compare Genomes/Phylogenetic Dist./Metagenomes vs. Genomes] using the “estimated gene copies” option (i.e., estimated by multiplying by read depth when available instead of using raw gene count) and the 60+ Perc. Identity option. A phylum-level table was downloaded for all samples. For Proteobacteria, a class-level table was also downloaded and merged, composing a table with all phyla and Proteobacteria divided into their classes. Estimated gene copies for each sample were divided by the total copies in order to obtain relative abundances and correct for different sampling depths (Supplementary Table S4).

**Nucleocytoplasmic large DNA viruses (NCLDV).** Nucleocytoplasmic large DNA viruses (NCLDV) marker genes, including major capsid proteins and DNA polymerases, were detected in the 58 Malaspina deep metagenomes samples with the use of previously described procedures45 using NCVOC46 and PSI-BLAST (e-value <1e-3) (Supplementary Table S5). For the detection of virophages sequences, we first screened the metagenomic sequences by the proteome sequences of three virophages (Sputnik, Mavirus, OLV) using BLAST (e-value<0.001). Then the metagenomic hits were searched against UniRef100 using BLAST (e-value<1e-3). As a result, 365 metagenomic peptides had best hits to virophage sequences, of which 50 sequences exhibited >95% sequence identity to homologs from the Mavirus virophages infecting Cafeteria roenbergensis.

**Viral signal analysis.** The marker gene terl, large subunit of the terminase was used to assess the diversity of bacterial and archaeal viruses in the deep-sea microbial metagenomes. terl genes were identified in the proteins predicted from the 58 metagenomes through hits to the PFAM domains PF04466 (terminase_3), PF03237 (terminase_6), PF03354 (terminase_1), and PF05876 (terminase_Gpa). Genes shorter than 100 bp were discarded, leading to a dataset of 485 terl genes. For the family-level affiliation (i.e., Myoviridae, Podoviridae, or Siphoviridae) of these sequences was obtained from a best blastp hit to the RefSeqVirus database (threshold of 50 on bit score, 0.001 on e-value, and 50% on amino acid identity; Supplementary Table S6). For each sample, viral community composition was then calculated based on this family-level affiliation and the normalized coverage of the contig (i.e., contig coverage divided by contig length and sequencing depth of the sample, as in Brum et al.46). Next, these 485 “deep-sea” terl genes were clustered with all terl from the RefSeqVirus database (n = 899, v7, 09-2015), from
"environmental phages" in Genbank (n = 456, downloaded on 07-2015), from the VirSorter Curated Database 96 (n = 6600), and from the Global Ocean Virome Dataset 96 (n = 2674, sequences from 91 Epi- and Mesoplagic viromes from Tara Oceans and Malaspina expedition) at 98% of nucleotide identity (threshold most consistent with genome-based population definition, i.e., ≥80% of genes shared at ≥95% average nucleotide identity, when tested on complete genomes from ReDoVirus), leading to 5701 OTUs. The 485 tefl genes from Malaspina were distributed across 14 phyla (Supplementary Table 2), including 38 unique to deep-sea samples (i.e., containing only Malaspina tefl sequences).

Marker enzymes from energy metabolisms. Metabolisms with a key role in the main biogeochemical cycles in the deep ocean were studied with additional detail by choosing specific enzymes for nitrogen, sulfur, methane, hydrogen, and carbon-fixation metabolic pathways. Marker enzymes for different pathways were defined by exploring the corresponding Kyoto Encyclopedia of Genes and Genomes (KEGG) 94 pathway maps (KEGG release 76.0). Enzymes participating only in reactions within each map were defined as marker enzymes for this metabolic path. When possible, marker enzymes were assigned to a specific module within a map (i.e., enzymes only participating in differentiation module within the nitrogen metabolism map). Corrected abundance estimation for marker enzymes was obtained by dividing the number of reads from the EC functional abundance table (without subsampling) by the number of reads assigned to the prokaryotic single-copy gene rrcA (selected as COG0668). A table selection of 83 KOs was built (Supplementary Data 3) representing the main key marker genes for different metabolic pathways relevant in the deep ocean. Of those, a total of 49 KOs were found in the Malaspina Deep-Sea Gene Collection (59%; Supplementary Table 8).

Nonmetric multidimensional scaling (NMDS) and permutational multivariate analysis of variance (PERMANOVA). Nonmetric multidimensional scaling (NMDS) with stable solution from random starts was performed for the ordination of samples in functional similarity using the KO abundance table, the Pfam, COG, and NC abundance tables (Supplementary Fig. 4). The Bray-Curtis distance measure was used for the abundance tables. All NMDS analyses were performed with the subsampled version of each abundance table. The partitioning of the variance in the Bray-Curtis distance matrix among oceans and oceanic basins was performed using permutational multivariate analysis of variance (PERMANOVA) 97 through the adonis function in the vegan R package (v2.5.6) 95 with 10,000 permutations. The two samples corresponding to the shallowest sampling station (station 62: 2400 m depth) were excluded from Fig. 2 as they appeared as clear outliers (i.e., showed the high distance to any other sample), but they were included in Supplementary Fig. 4.

Metagenomic Assembled Genomes from Deep Malaspina bathypelagic samples. To build the Malaspina Deep Metagenome-Assembled Genomes (MAGs) catalog (MDeep-MAGs), all 58 metagenomes from the Malaspina expedition were pooled and co-assembled (megahit v1.2.8; options:--preets meta-large-min-contig-len 2000) 96. Resultant contigs were de-replicated with cd-hit-est (v4.8.1) with a sequence similarity cut-off of 90%, with options -c 0.95 -n 10 -G 0 -as 0.95 -d 0 97. With this procedure, we increased the sequence space and we obtained a total of 421,891 contigs larger than 2000 bp. Metagenomic reads were back-mapped to the contigs dataset (bowtie 99 v2.3.4.1 with default options), keeping only mapping hits with quality larger than 10 (samtools 101 v1.8). We then assembled contigs into a total of 69 assemblies using the coverage and tetranucleotide frequencies in metatb (v2.12.1; igi summarize bam_contig_depths and metatb2 with default options) 111. The second round of assembly was carried out within each bin with CAP3 (v1.25-10-02; options -o 16 -p 95 -h 100 -f 9) to solve overlapping overhangs with 95% of sequence similarity between contigs. Assemblies from high-quality MAGs were examined visually in Geneious v10.2.4 and contigs that aligned fully to a larger contig were manually removed.

Taxonomic annotation of Deep Malaspina metagenomic assembled genomes (MDeep-MAGs). MDeep-MAGs’ completeness and single-copy gene redundancy (con- tamination) were estimated in CheckM (v1.0.18, lineage_wf) 105, and the placement in the tree of life for each MAG completeness larger than 50% and contamination lower than 10% was used to plot a tree depicting the phylogenetic relationships between them (Fig. 5; iTol 113 v4). Finer taxonomic assignations of the resulting 317 MAGs were estimated through the Genome Taxonomy Database (release r89) using GTDB-Tk (v1.0.2; classify_wf) (Supplementary Data 9). Briefly, the annotation relies on both taxonomic placement in a backbone tree (using marker genes) and average nucleotide identity (ANI) comparison based on ~150,000 genomes, spanning isolates, MAGs as well as SAGs.

Deep Malaspina MAGs annotation and metabolic prediction. All 619 bins were annotated, including gene prediction, rRNA, tRNA, and CRISPR detection with prokka (v1.13) with default options, using the estimated Domain classification from CheckM output as an argument of the --kingdom option. Additionally, MAG predicted coding sequences were annotated against the KEGG orthology database 14 with kofamscan 116 (v1.1.0; database timestamp 2019-10-15), and against the PfAM database (release 31.0) with hmmer 117 (v3.1.2B) with options--domtblout -E 0.01. Primarily KEGG orthology was used to determine the metabolism of the Malaspina Deep-Sea MAGs. A total of 83 marker genes from carbon fixation, methane, nitrogen, hydrogen, and sulfur metabolisms were selected (Supplementary Data 7) and their presence was explored along the low-quality (LQ) bins and medium quality (MQ) and high-quality (HQ) MAGs (https://metagenomics-public.github.io/malaspina-deep-ocean-microbiome/). A selected pool of 25 MAGs highlighting the potential for chemolithoautotrophy, mixotrophy, and non-cyanobacterial diazotrophs (NCDs) (Supplementary Data 9 and Supplementary Data 10) was further explored within the MDeep-MAGs.

Estimation of module pathway completeness within the Deep Malaspina MAGs. Module completeness per MAG was estimated by calculating the percentage of KO’s belonging to each module over its total KO number (KEGG release 2019-02-11).

Deep Malaspina MAGs abundance in the global bathypelagic Ocean. The abundance of each MAG was assessed by mapping competitively the reads from the 58 metagenomes against the MAGs contig database using blastn 104 (v2.7.1; options -perc_identity 70 -e_value 0.0001). Metagenomic reads were randomly subsampled to the smallest sequencing depth value (4,175,346 read pairs) with bowt0ol (v3.80, refomat.sh; https://sourceforge.net/projects/bbmap/) 118. Only reads with alignment coverage larger than 90% were kept for downstream analyses. Likewise, we kept only those metagenomic reads with sequence identity higher than 90%. In these cases, the same metabolic reads were assigned to the reference bin 119. In addition, we discarded reads mapping any region annotated as rRNA to avoid spurious hits to highly conserved regions. The abundance of each MAG was expressed as the number of mapped reads per genomic kilobase and sample gigabase.

Phylogenetic analyses of Rubisco, amoA, and nifH gene markers. The 18 Rubisco large-chain (K01461) amino acid sequences from the 619 Deep Malaspina bins were aligned using Clustal Omega 120 (v2.1.2/3 (default options and 100 iterations) against the Rubisco large-chain reference alignment profile published by Jaffe et al. 121, together with the large-chain sequences from heterotrophic marine Thaumarcheota published by Aylward and Santoro 122. Maximum-likelihood phylogenetic reconstruction was done with FastTree 223 (v2.1.11 (default options) using Jones-Taylor-Thorton model. Phylogenetic tree editing was done in iTol 113 v4.

For ammonia-oxidizing (amoA) gene phylogeny, we downloaded all nucleotide sequences annotated as amoA for phylum Nitrospirae and classes Beta and Gammaproteobacteria and Nitrosoospheeria, and all sequences annotated as pmoA for Archaea, Candidate division NC10, and classes Alpha and Gammamproteobacteria from NCBI (April 2020). We also included sequences of amoA of both marine epipelagic and deep pelagic Archaea from Alves et al 124 (clades NP-Alpha-2.2.1.1, NP-Epsilon-2 and NP-Gamma-2.1.3.1) to a total of 1651 sequence. We translated the sequences to amino acids (bacterial genetic code) and we removed redundancy by clustering all amino acid sequences to 98% of identity with cd-hit 125 (v4.8.1) to a final dataset of 586 sequences. We then added 19 sequences from 17 MAGs annotated with either K10944 or PF12942 and aligned all of them in mafft 126 (v7.402) with the G-INS-i option. We built a maximum-likelihood tree recording with FastTree 223 (v2.1.10) with default options and the Jones-Taylor-Thorton maximum-likelihood model and plotted it in iTol 113 v4. To analyze the phylogeny of the nifH gene of 3 potentially diazotrophic MAGs, we retrieved their blastp best hits against NCBI’s nr database (July 2020). We then added all nifH sequences between 240 and 320 residues long belonging to phylum Proteobacteria at NCBI Identical Protein Groups (IPG) database (query (nifH AND “240”[LEN]:”320”[LEN]))) AND proteobacteria[Organism]) and, additionally, we included 9 nifH sequences from MAGs from the Tara Oceans expedition 25, making a total of 192 sequences. Alignment, tree building, and plotting were done as above for amoA gene.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability. All data generated or analyzed during this study are included in this published article (and its supplementary information files). All raw data are publicly available at both DOE’s JGI Integrated Microbial Genomes and Microbiomes (IMG/MER) and the European Nucleotide Archive (ENA). Individual metagenome assemblies, annotation files, and alignment files can be accessed at IMG/MER. All accession numbers are listed in Supplementary Data 1. The co-assembly for the MAG dataset construction can be found through ENA at https://www.ebi.ac.uk/ena/ with accession number PRJEB40454. The nucleotide sequence for each MAG and their annotation files can be found through BioStudies at https://www.ebi.ac.uk/biostudies with accession S-BST457 and also in the companion website to this manuscript at https://malaspina-public.github.io/malaspina-deep-ocean-microbiome/.
Code availability
All software used in this work is publicly available distributed by their respective developers, and it is described in "Methods", including the versions and options used. Additional custom scripts to assign taxonomy to the M-geneDB genes and to filter and format FRA results are available through BioStudies at https://www.ebi.ac.uk/bioStudies with accession S-BST457.

Received: 25 September 2020; Accepted: 16 April 2021; Published online: 21 May 2021

References
1. Cho, B. C. & Azam, F. Major role of bacteria in biogeochemical fluxes in the ocean’s interior. Nature 332, 441–443 (1988).
2. Bar-On, Y. M., Phillips, R. & Milo, R. The biomass distribution on Earth. Proc. Natl Acad. Sci. USA 115, 6506–6511 (2018).
3. Aristegui, J., Gasol, J. M., Duarte, C. M. & Herndl, G. J. Microbial oceanography of the dark ocean’s pelagic realm. Limnol. Oceanogr. 54, 1501–1529 (2009).
4. Baltar, F., Aristegui, J., Gasol, J. M., Lekunberri, I. & Herndl, G. J. Mesoscale eddies: hotspots of prokaryotic activity and differential community structure in the ocean. ISME J. 4, 975–988 (2010).
5. Baltar, F. et al. Significance of non-sinking particulate organic carbon and dark CO2 fixation to heterotrophic carbon demand in the mesopelagic northeast Atlantic. Geophys. Res. Lett. 37, 109620 (2010).
6. Baltar, F. et al. Oceanography: dissolved organic carbon support of respiration in the dark ocean. Science 298, 1967 (2002).
7. Herndl, G. J. & Reinthaler, T. Microbial control of the dark end of the global carbon pump. Science 6, 718–724 (2013).
8. Baltar, F. et al. Significance of non-sinking particulate organic carbon and dark CO2 fixation to heterotrophic carbon demand in the mesopelagic northeast Atlantic. Geophys. Res. Lett. 37, 109620 (2010).
9. Boyd, P. W., Claustre, H., Levy, M., Siegel, D. A. & Weber, T. Multi-faceted particle pumps drive carbon sequestration in the ocean. Nature 568, 327–335 (2019).
10. Stukel, M. R., Song, H., Goericke, R. & Miller, A. J. The role of subduction and remineralization length scale in the California Current Ecosystem. Limnol. Oceanogr. 63, 363–383 (2018).
11. Omand, M. M. et al. Eddy-driven subduction exports particulate organic carbon from the spring bloom. Science 348, 222–225 (2015).
12. Jónasdóttir, S. H., Visser, A. W., Richardson, K. & Heath, M. R. Seasonal copepod lipid pump promotes carbon sequestration in the deep North Atlantic. Proc. Natl Acad. Sci. USA 112, 12122–12126 (2015).
13. Dall’Olmo, G., Dingle, J., Polimene, L., Brewin, R. J. W. & Claustre, H. Substantial export of organic carbon to the mesopelagic ecosystem from the seasonal mixed-layer pump. Nat. Geosci. 9, 820–823 (2016).
14. Herndl, G. J. et al. Contribution of archaea to total prokaryotic production in the deep Atlantic Ocean. Appl. Environ. Microbiol. 71, 2303–2309 (2005).
15. Wuchter, C. et al. Archaeal nitrification in the ocean. Proc. Natl Acad. Sci. USA 103, 12317–12322 (2006).
16. Reinthaler, T., van Aken, H. M. & Herndl, G. J. Major contribution of autotrophy to microbial carbon cycling in the deep North Atlantic’s interior. Deep Res. Part II Top. Stud. Oceanogr. 57, 1572–1580 (2010).
17. Swan, B. K. et al. Potential for chemolithoautotrophic growth among ubiquitous bacteria lineages in the deep ocean. Science 333, 1296–1300 (2011).
18. Pachiadaki, M. G. et al. Major role of nitrite-oxidizing bacteria in dark ocean carbon fixation. Science 358, 1046–1051 (2017).
19. Högler, M. & Sievert, S. M. Beyond the Calvin Cycle: autotrophic carbon injection by the biological pump. Nat. Commun. 6, 7608 (2015).
20. Smith, K. L., Bubl, H. A., Hulfard, C. L., Messé, M. & Kahru, M. Episodic organic carbon fluxes from surface ocean to abyssal depths during long-term monitoring in NE Pacific. Proc. Natl Acad. Sci. USA 115, 12235–12240 (2018).
21. Boeuf, D. et al. Biological composition and microbial dynamics of sinking particulate organic matter at abyssal depths in the oligotrophic open ocean. Proc. Natl Acad. Sci. USA 116, 11824–11832 (2019).
22. Mestre, M. et al. Sinking particles promote vertical connectivity in the ocean’s interior. Science 311, 496–503 (2006).
23. Salm, A. et al. Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. Cell 179, 1068–1083, e21 (2019).
24. Baltar, F. et al. Prokaryotic extracellular enzymatic activity in relation to biomass production and respiration in the meso- and bathypelagic waters of the (sub)tropical Atlantic. Environ. Microbiol. 11, 1998–2014 (2009).
25. Bergauer, K. et al. Organic particle remineralization by microbial communities throughout the Atlantic water column as revealed by metaproteomics. Proc. Natl Acad. Sci. USA 115, E400–E408 (2018).
26. Salazar, G. et al. Global diversity and biogeography of deep-sea pelagic prokaryotes. ISME J. 10, 598–608 (2016).
27. Anantharaman, K., Breier, J. A., Sheik, C. S. & Dick, G. J. Evidence for substrate turnover by attached bacteria. Limnol. Oceanogr. 53, 1878–1886 (2008).
28. Agusti, S. et al. Ubiquitous healthy diatoms in the deep sea confirm deep carbon injection by the biological pump. Nat. Commun. 6, 7608 (2015).
29. Ruiz, M. et al. Physiological, ecological, and phylogenetic characterization of Stappia, a marine CO2-oxidizing bacterial genus. Appl. Environ. Microbiol. 73, 1266–1276 (2007).
30. Martin-Cuadrado, A. B., Ghai, R., Gonzaga, A. & Rodríguez-Valera, F. CO dehydrogenase genes found in metagenomic fosmid clones from the deep Mediterranean Sea. Appl. Environ. Microbiol. 75, 7436–7444 (2009).
Acknowledgements

We thank the R/V Hespérides captain and crew, the chief scientists in Malaspina expedition legs, and all project participants for their help in making this project possible. This work was funded by the Spanish Ministry of Economy and Competitiveness (MINECO) through the Consolider-Ingenio program (Malaspina 2010 Expedition, ref. CSD2008-00077). The sequencing of 58 bathypelagic metagenomes was done by the U.S. Department of Energy Joint Genome Institute, supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02 05CH11231 to SGA (CSP 612 – Metagenomics reveal. Mol. Biol. Evol. 36, 539 (2011).

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-021-02112-2.

Correspondence and requests for materials should be addressed to S.G.A.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.