scientific reports

Check for updates

OPEN

The influence of depth on the global deep-sea plasmidome

Melany Calderón-Osorno[1,2,7]✉, Dorian Rojas-Villalta[1], Franck Lejzerowicz[3], Jorge Cortés[4,5], Maria Arias-Andres[6] & Keilor Rojas-Jimenez[5,7]✉

Plasmids play a crucial role in facilitating genetic exchange and enhancing the adaptability of microbial communities. Despite their importance, environmental plasmids remain understudied, particularly those in fragile and underexplored ecosystems such as the deep-sea. In this paper we implemented a bioinformatics pipeline to study the composition, diversity, and functional attributes of plasmid communities (plasmidome) in 81 deep-sea metagenomes from the Tara and Malaspina expeditions, sampled from the Pacific, Atlantic, and Indian Oceans at depths ranging from 270 to 4005 m. We observed an association between depth and plasmid traits, with the 270–1000 m range (mesopelagic samples) exhibiting the highest number of plasmids and the largest plasmid sizes. Plasmids of Alphaproteobacteria and Gammaproteobacteria were predominant across the oceans, particularly in this depth range, which also showed the highest species diversity and abundance of metabolic pathways, including aromatic compound degradation. Surprisingly, relatively few antibiotic resistance genes were found in the deep-sea ecosystem, with most being found in the mesopelagic layer. These included classes such as beta-lactamase, biocide resistance, and aminoglycosides. Our study also identified the MOBP and MOBQ relaxase families as prevalent across various taxonomic classes. This research underscores the importance of studying the plasmidome independently from the chromosomal context. Our limited understanding of the deep-sea's microbial ecology, especially its plasmidome, necessitates caution in human activities like mining. Such activities could have unforeseen impacts on this largely unexplored ecosystem.

Keywords Plasmids, Metagenomics, Deep-sea, Genetic exchange, Adaptive traits

Within microbial communities, the exchange of genetic information is primarily facilitated by mobile genetic elements (MGEs), especially plasmids[1,2]. Plasmids are extrachromosomal elements[3,4] containing a diverse array of genes that contribute to both their own functions and the physiology of their host organisms[5,6]. These genetic entities play a pivotal role in carrying genes crucial for adaptive evolution, thereby enhancing the fitness of the host organism[7]. Different environmental conditions contribute to distinctive bacterial community compositions, potentially impacting the predominance of specific associated plasmids, as it is generally assumed that only a subset of the bacterial community carries plasmids[8]. Understanding the associated plasmids, the types of traits they carry, and the reasons behind their presence remains an open question in plasmid ecology[8,9].

As of 2024, the Plasmid Database (PLSDB) from the National Center for Biotechnology Information (NCBI) contains approximately 59,800 deposited plasmids, with 25,382 new additions since the previous version. However, marine plasmids remain notably underrepresented in this dataset. Studies indicate that less than 1% of the entries comprised marine plasmids, while other authors suggested that the proportion of marine plasmids is less than 2%[3,5]. These marine plasmids are typically isolated from diverse environments such as oceanic seawater, ocean sediment, marine biofilms, and marine mud. The lack of marine plasmid representation has led to significant knowledge gaps regarding the traits and their environmental impact[3,5]. The ongoing challenge of isolating plasmids persists, complicating efforts to expand data collection across all environments, including marine ecosystems[8,10,11]. Reconstructing plasmids from metagenomic data offers a solution to obtain draft plasmids[12,13], enabling their detection and characterization in marine environments[14]. Various strategies can be

[1]Costa Rica National High Technology Center (CeNAT), Pavas, San José 10108, Costa Rica. [2]Maestría académica en Biología con enfásis en genética y biología molecular, University of Costa Rica, San Pedro, San José 11501-20260, Costa Rica. [3]Section for Aquatic Biology and Toxicology, University of Oslo, Blindernveien 31, 0371 Oslo, Norway. [4]Centro de Investigación en Ciencias del Mar y Limnología (CIMAR), University of Costa Rica, San Pedro, San José 11501-20260, Costa Rica. [5]Biology School, University of Costa Rica, San Pedro, San José 11501-20260, Costa Rica. [6]Central American Institute for Studies on Toxic Substances, Universidad Nacional, Campus Omar Dengo, Heredia 86-3000, Costa Rica. [7]Melany Calderón-Osorno and Keilor Rojas-Jimenez contributed equally to this work. ✉email: mcalderono@cenat.ac.cr; keilor.rojas@ucr.ac.cr

applied to analyze plasmids in metagenomic datasets, such as raw read profiling and contig assembly to identify plasmidic contigs[12]. This approach is particularly valuable given the limited number of annotated plasmids from oceanic environments within existing plasmid databases.

Although few studies have utilized this approach, reanalyzing marine metagenomic sequences offers an excellent opportunity to explore the study of plasmids in greater detail[1,14]. A study conducted in the Red Sea successfully detected 362 plasmid candidates from 45 samples, demonstrating that reanalyzing publicly available data can lead to the discovery and characterization of marine plasmids. The scarcity of studies in this field highlights the necessity for further research on plasmid detection in oceanic environments. This raises the question of how the population of plasmids varies along a depth gradient.

In this work, we characterize the plasmidomes of whole metagenomes from the deep-sea, defined as depths below 200 m, which play a crucial role in global biogeochemical cycles and nutrient recycling[15]. Once considered pristine, these ecosystems have been impacted by industrial activities, introducing microplastics and organic pollutants that affect marine life[16–19]. Furthermore, anthropogenic activities such as mining and microplastic pollution have the potential to disrupt these ecosystems, whose functions-particularly within the plasmidome-are only beginning to be understood[20,21].

The limited exploration and scientific research, due to the remote and harsh conditions of the deep ocean, have led to a poor understanding of the majority of this vast ecosystem[20]. Biological communities in this environment are expected to recover slowly following disturbances[20]. Despite efforts to monitor deep-sea disturbances, there is a lack of research on the distribution of traits in deep-sea plasmidomes[22,23]. In this ecosystem, the contributions of the microbiome and plasmidome could be of great importance. However, these elements have not been previously characterized, making this study the first to address this gap in knowledge.

Our analysis focuses on plasmidomes derived from 81 metagenomes collected from two distinct depth groups. To achieve this, we employ a bioinformatics pipeline that includes tools for extracting plasmid sequences, taxonomic annotation, identification of resistance and mobility genes, and pathway annotation.

Results and discussion

Profiling

In this study, we reanalyzed 81 whole metagenomic sequences obtained from the Atlantic, Indian, and Pacific Oceans[24,25]. To our knowledge, this is the first comprehensive exploration of the deep-sea plasmidome on a global scale. Previous studies[24,26–29] have focused on broader aspects of bathypelagic microbial communities without explicitly addressing plasmids. These earlier works have highlighted key drivers of microbial structure and functionality, particularly among free-living and particle-attached prokaryotes. For example, Acinas et al.[24] demonstrated that distinct lifestyles shape metabolic pathways, with ammonia and CO oxidation dominating free-living communities, while particle-attached groups are enriched in dissimilatory nitrate reduction and H2 oxidation pathways. The Calvin-Benson-Bassham cycle was identified as the primary carbon fixation pathway in both lifestyles, with mixotrophy emerging as a significant ecological trait.

Rigonato et al.[26] linked oxygen, nitrate, and particle flux to mesopelagic community composition, identifying three key modules: oxic-associated prokaryotes and pico-eukaryotes, OMZ-enriched prokaryotes, and OMZ pico-eukaryotes. Salazar et al.[27] further underscored the role of water masses in shaping biogeographical patterns of particle-attached microbes, driven by water mass aging and limited dispersal. Meanwhile, Sebastián et al.[28] and Sunagawa et al.[29] elucidated the ecological roles of genes like *merA* and *merB*, involved in mercury cycling, and revealed shared core functionalities between oceanic and human gut microbiomes.

Our study bridges a critical gap by focusing on plasmids as genetic elements driving functional diversity and ecological adaptation in deep-sea microbial communities. Through plasmid-focused analyses, we complement established knowledge on the roles of oxygen, particle flux, and water masses.

For raw reads profiling, plasmid sequence communities across all oceans were predominantly associated with hosts assigned to Alphaproteobacteria and Gammaproteobacteria. Specifically, Alphaproteobacteria ranged from 24.4% to 74.3% of the reads, with an average of 53.4%, while Gammaproteobacteria ranged from 2.7% to 63.5%, with an average of 25.3%. Actinomycetes constituted the third most prevalent group of plasmid hosts, ranging from 1.19% to 17.2%, with an average of 6.1%. Notably, Actinomycetes-associated plasmids dominated in the Indian Sea (Fig. 1). Chromosome sequence communities predominantly were associated to Alphaproteobacteria and Gammaproteobacteria across all oceans. Specifically, Alphaproteobacteria were more prevalent in the Indian Sea, whereas Gammaproteobacteria dominated the Atlantic Ocean (Fig. 1).

Across all oceans, the depth range of 270 to 1000 m (mesopelagic) exhibited the highest richness and diversity of species detected based on chromosome and plasmid reads, averaging 1300 species on chromosomes and 101 host species on plasmids. Depths ranging from 2400 to 4005 m (bathypelagic) showed a lower number of species, with an average of 588 species on chromosome reads and 52 species on plasmid reads. Chromosome diversity decreased across depth group, with a mean Shannon index of 4.1 in mesopelagic samples and 3.8 in bathypelagic samples (2400-4005 m). However, plasmid diversity appeared to remain stable across depths, with a consistent mean Shannon index of 3.1, suggesting that plasmid-associated species are less affected by increasing depth. Statistical analysis revealed significant differences among depths for chromosome diversity (p-value = 1.7e-4) and richness (p-value = 3.5e-11), as well as for plasmid richness (p-value = 8.6e-8).

In the 270-1000 m depth group, the Atlantic Ocean exhibits an average of 1553 species (Fig. 2a) and a mean Shannon index of 4.1 (Fig. S1a) for species detected based on chromosomes. In bathypelagic samples, the Pacific Ocean shows a larger richness, with an average of 624 species (Fig. 2a), while the Indian Ocean showed the highest mean Shannon index of 3.9 (Fig. S1a) among chromosome derived species. The plasmidic richness and diversity remained more stable across depths. The Atlantic Ocean presented the highest plasmidic richness in both depth groups, with an average of 121 species (Fig. 2b) at the 270-1000 m depth and an average of 52

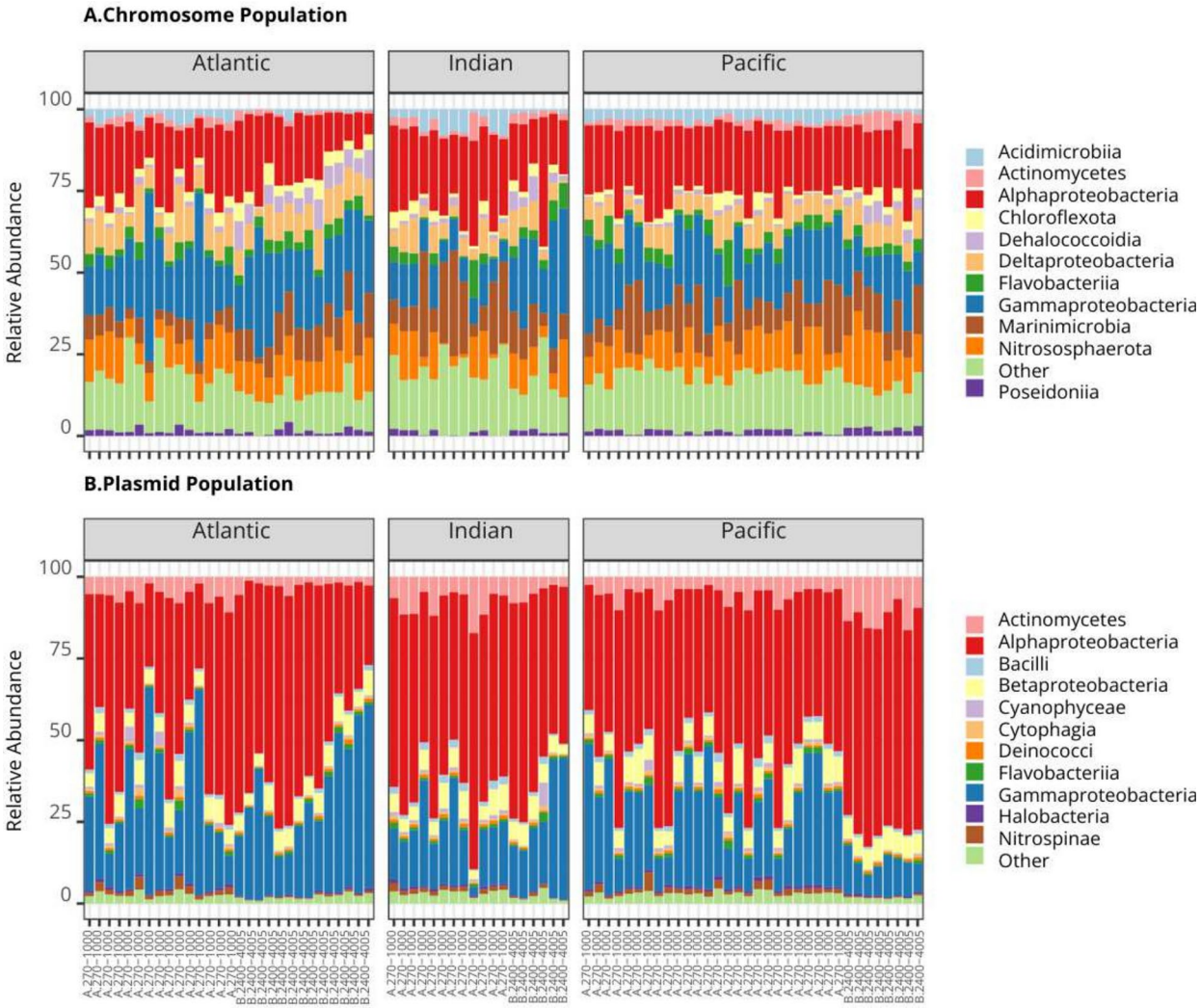

Fig. 1. Relative abundance of the main taxonomic groups in deep-sea samples grouped by depth for each ocean. Panel A shows the taxonomic composition based on predicted chromosomal DNA, while Panel B shows the composition based on plasmids. Classes that are shared between chromosomal DNA and plasmids are represented with the same color.

species (Fig. 2b) at the 2400-4005 m depth. The Pacific Ocean shows greater diversity for species associated with plasmids across both depths, with a mean Shannon index of 3.2 (Fig. S1b).

Although many forces can shape plasmid communities and their interplay with the hosts, studies carried out in diverse environments report that plasmidome composition could be more influenced by stochasticity[30]. Here, we identified that depth has a strong influence on the host microbial community and observed that the main groups identified in plasmid diversity were also found in chromosome populations.

A permutational multivariate analysis of variance (PERMANOVA) was conducted on Bray-Curtis dissimilarities to assess the impact of oceans and depth groups on the composition of chromosome and plasmid microbial communities. The analysis revealed that both oceans (R^2 = 0.11) and depth groups (R^2 = 0.22) significantly contribute to the dissimilarity in chromosome microbial communities (p-value = 0.001). Similarly, in plasmid communities, oceans (R^2 = 0.08) and depth groups (R^2 = 0.17) were significant factors (p-value = 0.001). However, in plasmidomes, a considerable portion of the variance remains unexplained, even when other factors such as latitude and longitude are considered. Visualization of Bray-Curtis dissimilarities using a non-metric multidimensional scaling (NMDS) plot showed that samples clustered according to depth groups (Fig. 2c,d), for chromosomal and plasmid communities, which is consistent with the results shown for the composition and also the alpha diversity values.

Plasmidomes annotation and plasmid sequence contigs profiling

Following read assembly into contigs and plasmidic contig prediction, a total of 10,181 plasmid sequence contigs were identified from 29,687,283 assembled contigs (approximately 0.03%) using heuristic parameters by Platon[31]. Among these potential plasmid contigs, 787 (7.7%) were annotated as circular by at least one of three circularity detection methods: "ccfind," "Platon", or the "reverse-forward" approach[31,32] (Fig. S2). Although the

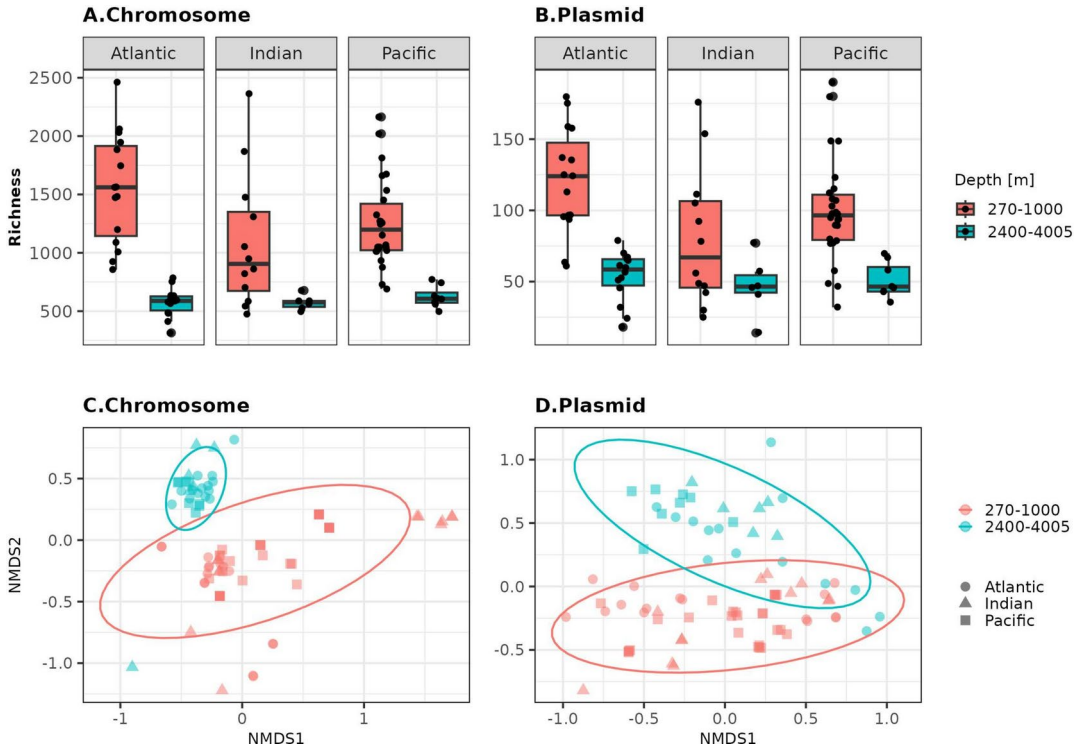

Fig. 2. Richness of microbial communities in chromosomes (**A**) and plasmid reads in the deep-sea across the three oceans (**B**). The boxplots (**A**,**B**) depict the distribution of data within Ocean and depth group categories. The non-metric multidimensional scaling (NMDS) plot represents the microbial community using Bray-Curtis dissimilarity for chromosomes (**C**) and plasmids (**D**) reads. Points that are closer together indicate a higher similarity in community composition, while points that are farther apart indicate a lower similarity in community composition.

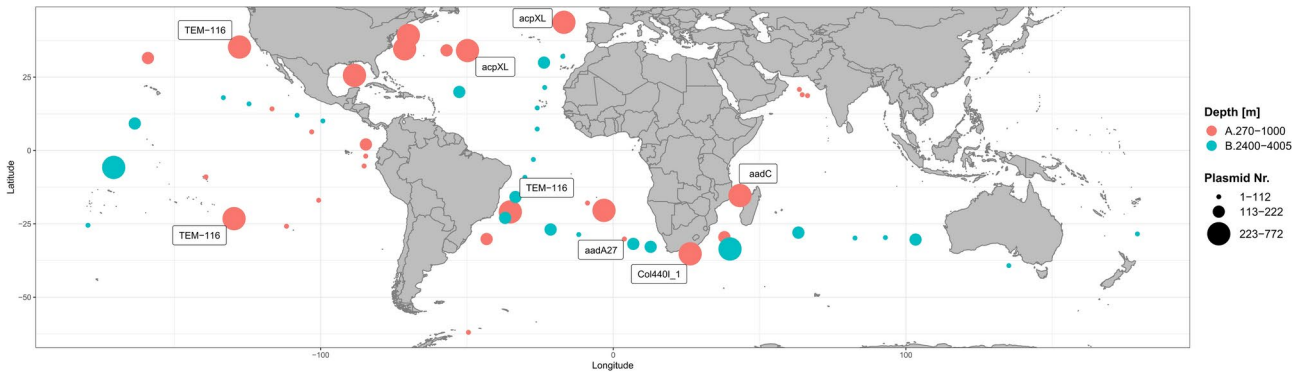

Fig. 3. Map displaying the global Plasmidome in the deep-sea samples analyzed in this study. Dot size is based on the number of plasmid sequence contigs identified, and the color indicates the depth group to which they belong. Samples containing annotated antibiotic resistance genes are highlighted.

isolated DNA was not specifically enriched for plasmids, we identified 10,181 plasmid candidates. It is worth noting that plasmid detection typically requires dedicated plasmid extraction protocols to achieve optimal recovery[3]. Furthermore, the dataset used in this study was derived from the 0.2-0.8 μm fraction, which excludes most eukaryotes, such as diatoms, known to harbor plasmids[24,25]. Therefore, these plasmid candidates likely represent only the tip of the iceberg of the deep-sea plasmidome.

The distribution of plasmid contigs across the oceans was approximately 40.9% in the Atlantic Ocean, 24.0% in the Indian Ocean, and 35.1% in the Pacific Ocean. A higher number of plasmid contigs was observed in mesopelagic samples (270-1000 m), with a total of 2,868, 2,631 and 1,702 plasmid contigs from the Pacific, Atlantic and Indian Oceans, respectively (Fig. 3, Fig. S3).

When plasmid contigs were evaluated using common descriptive metrics (total number of contigs produced, N50, total assembly size, and maximum contig length), all metrics varied among the depth groups, with the best plasmid contig assembly observed at 270-1000 m (mesopelagic samples) (Fig. S4). At this depth, an average of 2,631 plasmid contigs were identified (Fig. S4a), with the highest total assembly size (average of 16 Mb) (Fig. S4c) and the largest contig size (average of 2 Mb) (Fig. S4d). Assembly quality was significantly affected by depth (Kruskal-Wallis test, p-value = 0.04) but not by ocean (Kruskal-Wallis test, p-value = 0.9).

Additionally, plasmid contig size exhibited significant variation across depth groups (p-value = 2.5e-5, Kruskal-Wallis test), with plasmid sequence contigs being larger in mesopelagic samples (270-1000 m) (average of 52 kb) compared to bathypelagic samples (2400-4005 m) (average of 8.9 kb) (Fig. S5). Oceanic differences did not significantly affect plasmid contig size (p-value = 0.5, Kruskal-Wallis test). Among circular plasmidic contigs, the majority (83.1%) were identified in mesopelagic samples. Depth significantly influenced both the size (Fig. S6) and RPKM values (Fig. S7) of these contigs. Interestingly, circular plasmid contigs in bathypelagic samples were larger and exhibited higher RPKM values compared to those in mesopelagic samples (Fig. S6, Fig. S7). The observed average plasmid contig sizes aligns with typical sizes reported for both mobilizable and non-transmissible plasmids, which have been documented to range from 4 to 100 kb[1,19,33–35].

Plasmid sequence contigs communities were predominantly associated with hosts assigned to Alphaproteobacteria (ranging from 0% to 100%, with an average of 48.2%) and Gammaproteobacteria (ranging from 0% to 100%, with an average of 41.3%) across all oceans, with Actinomycetes being the third most prevalent class (ranging from 0% to 66.3%, with an average of 6.7%). Alphaproteobacteria were more dominant in the Pacific Ocean (ranging from 0.6% to 100%, average of 52.8%), while Gammaproteobacteria were abundant within the Atlantic Ocean (ranging from 0% to 100%, average of 47.5%). Notably, Actinomycetes were dominant in the Indian Sea (ranging from 0% to 63.6%, with an average of 15.6%) (Fig. S8), which aligns with the results obtained from raw read profiling.

Our results on the prevalence of plasmids taxonomy align with previous studies from the analyzed datasets, which indicate that Gammaproteobacteria and Alphaproteobacteria were the most abundant classes[24,27]. It has been described that Alphaproteobacteria and Gammaproteobacteria dominate marine water ecosystems[36,37].

We determined no statistically significant differences in the taxonomic richness and diversity of plasmid sequence contigs when comparing different depth ranges. In both depth groups, the Atlantic Ocean exhibited the highest richness values (44 and 25 species for depths of 270-1000 and 2400-4005 m, respectively). The Shannon index values were higher in the Atlantic Ocean (average = 2.2) for mesopelagic samples (270-1000 m), while for bathypelagic samples (2400-4005 m), the higher diversity was determined in the Pacific Ocean with an average Shannon index of 2.1 (Fig. S9). PERMANOVA analysis of plasmid sequence contigs revealed a significant effect of Ocean (R2 = 0.05, p = 0.005) and depth group (R2 = 0.12, p = 0.001) on the dissimilarity of the communities. The NMDS plot illustrates that samples tend to cluster according to depth groups (Fig. S10). These taxonomic and alpha diversity measures reaffirm the dominance of a small group of highly abundant hosts in shaping plasmid community composition within oceans.

Functional annotation and antibiotic resistance genes (ARGs) mining

Functional protein sequences from annotated plasmid contigs were subjected to analysis using GhostKOALA[38] to identify metabolic pathways (Fig. 4). A total of 36 complete metabolic pathways were identified, encompassing various processes such as nitrate assimilation, citrate oxidation, amino acid biosynthesis (e.g., cysteine), respiratory activities involving cytochrome complexes, carbohydrate degradation and synthesis, lipid and coenzyme biosynthesis, and antibiotic resistance mechanisms.

This finding is consistent with previous reports indicating that such processes are expected to be commonly found in most plasmids[1,39–41]. Indeed, numerous studies have emphasized the importance of functions such as cell wall/membrane/envelope biogenesis, energy production and conversion, post-translational modifications and protein turnover, carbohydrate metabolism and transport, amino acid transport and metabolism, inorganic ion transport and metabolism, intracellular trafficking, secretion, and vesicular transport in the functional profile crucial for plasmid maintenance and survival (including replication, conjugation, and resistance)[1].

Notably, in the Atlantic Ocean, the most abundant metabolic pathways were aromatic compound degradation (present in 14.3% of the samples), fatty acid degradation (57.1% of the samples), and central carbohydrate metabolism (35.7% of the samples). In the Indian Ocean, predominant pathways included nucleotide biosynthesis (14.3% of the samples), amino acid metabolism (14.3% of the samples), fatty acid degradation (85.7% of the samples), and carbohydrate metabolism (57.1% of the samples). Similarly, in the Pacific Ocean, fatty acid degradation (present in 62.5% of the samples) and carbohydrate metabolism (43.8% of the samples) were the most prevalent pathways.

Studies have highlighted the capability of many Gammaproteobacteria members, which were most prevalent in the Atlantic Ocean, to degrade aromatic compounds[42]. Notably, authors like Gutierrez[43] have emphasized that the Gammaproteobacteria class harbors the most significant genera and highest diversity of obligate and generalist hydrocarbonoclastic bacteria in marine environments. Genera such as *Acinetobacter*, *Alcanivorax*, *Alteromonas*, *Halomonas*, *Marinobacter*, *Paraburkholderia*, and *Pseudoalteromonas*, identified in plasmid sequence contigs (Fig. S11), are known hydrocarbonoclastic bacteria primarily confined to marine environments, where they are commonly enriched at oil-contaminated sites[44]. According to The International Tanker Owners Pollution Federation Limited (TOPF), the Atlantic Ocean has the regions with the highest concentration of oil spills[45]. Microplastic contamination has also been reported in the Atlantic Ocean[46]. In the Pacific Ocean, where Alphaproteobacteria were predominantly identified in plasmid contigs profiling, prevalent pathways included carbohydrate metabolism and fatty acid degradation. These pathways are integral components of plasmids[1,26,40,41].

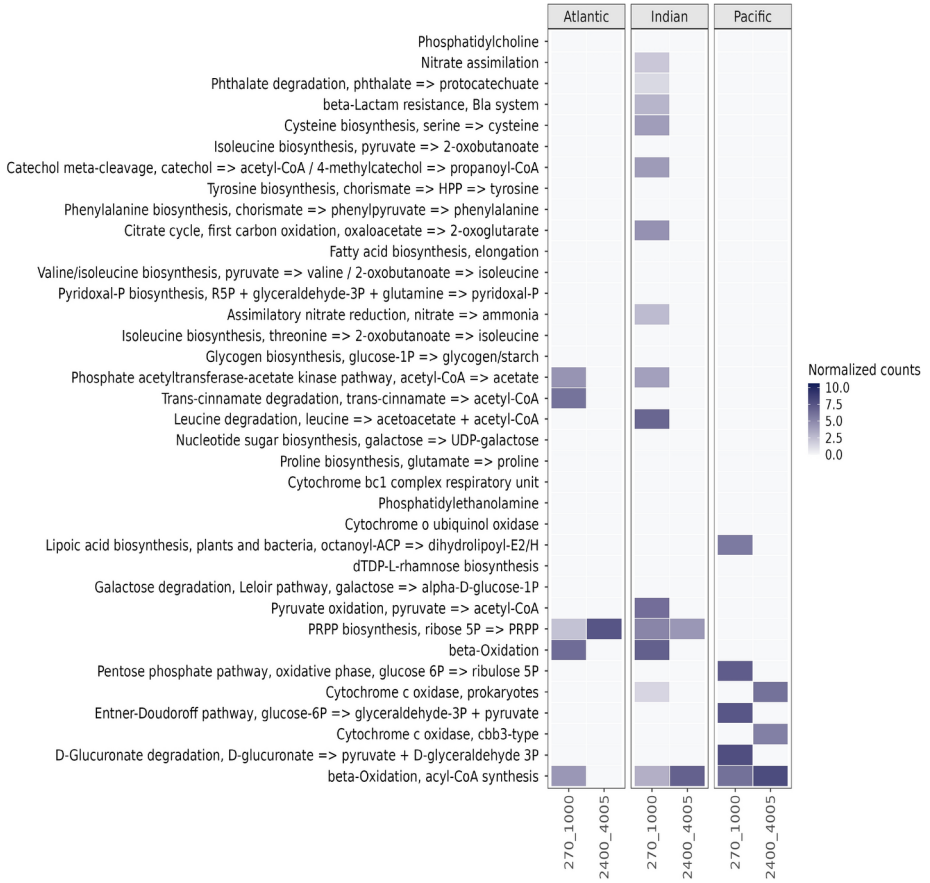

Fig. 4. Heatmap of the presence of complete KEGG modules in plasmid sequence contigs, across oceans and grouped by depth. The scale of colors indicates the abundance of the KEGG modules. The value of 10 represents the higher KEGG module abundance[38].

Interestingly, the 270-1000 m depth group (mesopelagic) in all three oceans exhibited the highest abundance and diversity of pathways, with beta-oxidation and acyl-CoA synthesis being the most prevalent (Fig. 4). One plausible hypothesis is that shallower depths are subject to greater environmental stress[44]. It is well established that plasmids frequently carry a suite of accessory genes that confer various ecological traits upon their bacterial hosts, including resistance to stressors and diverse metabolic capabilities[7]. These traits may be more required at superficial depths. Notably, the most prevalent pathway observed within mesopelagic samples was a crucial component of the fatty acid degradation mechanism.

Notably, we detected five antibiotic resistance genes across all samples. The analysis revealed the presence of genes associated with beta-lactamase, biocide resistance, and aminoglycoside classes. Notably, antibiotic-resistance genes were predominantly found in mesopelagic samples (270-1000 m) (Fig. S12a), with aminoglycoside resistance being the most prevalent class in the Atlantic Ocean (average of 3.5 normalized counts) and beta-lactamase resistance prevailing in the Pacific Ocean (average of 8.7 normalized counts) (Fig. S12). Few samples were identified that carry ARGs on plasmids (Fig. S12b). Samples from the Atlantic Ocean contained the *aadA27*, *ANT3-PRIME*, and *TEM-116* resistance genes, while samples from the Pacific Ocean contained only the TEM-116 gene (Fig. S12b).

The minimal presence of antibiotic resistance genes in the fragile deep-sea ecosystem is promising news. However, the persistence of this situation is uncertain due to the rising detection of ARGs in polluted rivers and coastal marine areas. Studies have shown that rivers and coastal environments are increasingly being contaminated with ARGs, largely due to wastewater discharge and pollution from human activities. This trend poses significant environmental and public health risks, indicating that the relatively low levels of ARGs in deep-sea environments may not remain stable for long[47].

Recent research has also highlighted the presence of various antibiotic classes, including sulfonamides, tetracyclines, beta-lactams, macrolides, and aminoglycosides, in marine microorganisms[48]. Within our study, resistance genes were predominantly detected in the 270-1000 m depth group, with aminoglycoside resistance being the most prevalent class in the Atlantic Ocean and beta-lactamase resistance prevailing in the Pacific Ocean (Fig. S13). Moreover, a recent investigation into deep-sea water (deeper than 1000 m) suggests that even these remote marine environments could serve as reservoirs for antibiotic resistance genes, primarily carried by representatives of Gammaproteobacteria (70%) and Alphaproteobacteria (20%)[49].

To gather indirect evidence of the main conjugation and HGT mechanisms of the plasmids, we analyzed the presence of MOB family relaxases, conjugative proteins that initiate bacterial conjugation by a site-specific cleavage of the transferred DNA strand[50]. We identified six different MOB families of relaxases in plasmid sequence contigs from deep-sea waters. The most prevalent families associated with plasmid mobility were MOBP and MOBQ (Fig. S14a). MOBP (padj = 1.5e-3, 4.8 log2 fold change) and MOBQ (padj = 2.3e-2, 4.2 log2 fold change) were predominant in the 2400-4005 m depth group (mesopelagic samples), with MOBP being most prevalent in the Pacific Ocean (average of 11.2 normalized counts) and MOBQ being prevalent in the Atlantic Ocean (average of 6.9 normalized counts) at this depth (Fig. S14a). Furthermore, the composition of MOB genes varied depending on the associated host phylum, with MOBP predominantly found in contigs annotated to the Alphaproteobacteria and MOBQ more abundant in Gammaproteobacteria (Fig. S14b). The MOB families of relaxases were characterized by the predominance of MOBQ and MOBP. Specifically, MOBQ with an average of 30.3%, while MOBP with an average of 29.1%, across all oceans. MOBF constituted the third most prevalent family, with an average of 27.1% (Fig. S15b).

The set of mobility (MOB) genes is essential for conjugative DNA processing. These genes, also known as Dtr genes (for DNA-transfer replication), facilitate the transfer of DNA[50].This finding aligns with previous research that has reported the presence of MOBP, MOBF, MOBQ, MOBH, MOBV, and MOBC families, with MOBP being the most abundant in marine environments[6].

Conclusions

Our findings revealed that various plasmid traits, including size, community diversity, antibiotic resistance and mobility genes, and pathways, exhibit notable variations according to depth, being higher in mesopelagic samples compared to bathypelagic samples across all oceans. Despite the inherent challenge of extracting useful information from short sequence reads derived from chromosome data[12], we obtained draft plasmid sequence contigs that were subsequently characterized.

Overall, our study sheds light on the intricate dynamics of the deep-sea plasmidome, providing valuable insights into its composition, diversity, and functional attributes. Our reanalysis of metagenomic sequences from the Atlantic, Indian, and Pacific Oceans has provided valuable insights into the composition of plasmid communities within diverse marine environments. We observed a strong association between sampling depth and the characteristics of detected plasmids, with the 270-1000 m (mesopelagic) exhibiting the highest abundance and diversity of plasmids, alongside the largest plasmid contig sizes. Moreover, this depth range showed a higher prevalence of antibiotic resistance genes.

Additionally, the dominance of specific taxa such as Alphaproteobacteria and Gammaproteobacteria, along with the prevalence of metabolic pathways like aromatic compound degradation and fatty acid metabolism at these depths, underscores the ecological significance of marine layers in shaping plasmid community composition and functional diversity. These findings deepen our understanding of plasmid-mediated gene transfer dynamics and their broader implications for microbial ecology and biogeochemical processes in ocean ecosystems. One limitation of our study is that depth groups were established based on the sampling depth of each study. Future endeavors could address this by using bioinformatics tools such as Simka to assess how the samples diverge as the studies vary significantly through subsampling. Simka enables k-mer subsampling and calculates diversity metrics based on randomly selected k-mers[51,52]. This approach allows us to assess whether depth continues to influence diversity, even when the samples are subsampled.

Methods

Data acquisition

In order to analyze the global deep-sea plasmidome, we searched for raw reads of deep-sea shotgun sequencing samples of seawater from around the globe in the European Bioinformatics Institute (EBI) European Nucleotide Archive (ENA) database[53] and NCBI databases[54]. Two main expedition bioprojects were selected for our study: the Malaspina Expedition 2010 (accession number PRJEB40454)[24] and the Tara Oceans Expedition 2009-2013 (accession number PRJEB1787)[25]. Each expedition provided samples collected from distinct depths, with the Tara Oceans expedition covering depths from 270 to 1000 m and the Malaspina expedition covering depths from 2400 to 4005 m. We retrieved a total of 81 whole metagenomes samples from the Indian Ocean (n = 18), Atlantic Ocean (n = 29) and Pacific Ocean (n = 34) (accession date: September 27, 2023) The sample accession numbers, locations and depths, are shown in Supplementary Table S1.

The sample identifiers and associated metadata were acquired from the ENA database[53]. The samples were collected at various sites and depths. We established a variable termed "depth group," categorizing samples based on their depth into "270_1000" for mesopelagic samples and "2400_4005" for bathypelagic samples. These categories correspond to the depth ranges explored during each expedition. A table specifying each biosample accession and other metadata can be found in Table S1.

Data processing and profiling

The quality of the raw reads was assessed using FastQC v0.11.9[55]. We trimmed the sequences utilizing fastp v0.20.1[56] with a minimum average quality value of 25, followed by filtering for human contamination using Bowtie2 v2.5.0[57]. The filtered reads were taxonomically classified at species levels using the GTDB reference database v207[58] with Kraken2 v2.1.2[59] and Bracken v2.7[60]. This classification process was performed separately for sub-databases corresponding to chromosomal and plasmid sequences of the GTDB species records. While plasmid raw reads could be associated with several species, taxonomy was assigned to the taxon with the highest similarity. The two species classifications were then utilized to construct feature tables expressing read counts per sample and taxonomy.

Whole metagenome assembly and plasmidomes annotation

Filtered and processed reads were used for assembly with metaSPAdes v3.14.1[61], with k-mer values of 33, 55, 77 and 99 bp. Assembled contigs shorter than 1000 bp were filtered through the reformat.sh function from bbmap v37.36[62]. We employed Platon v1.6[31] and PlasX[32] to extract plasmid sequence contigs from the whole metagenomes, as Platon has shown the best performance at retrieving a high number of potential plasmids from aquatic metagenomes among AI-based tools[63] and PlasX was recently proposed as further improving detection with a scoring scheme based on modeling plasmid gene family frequencies. Contig circularity was identified using "ccfind"[64] and by detecting "reverse-forward" oriented reads paired over contig ends. This latter approach identified paired reads mapped in reverse-forward orientation, forming a gap longer than the contig length minus three times the median insert size between forward-reverse oriented read pairs (see Method in[32]). It is important to consider that due to sequencing technology limitations, two contigs that actually belong to the same plasmid might not be assembled correctly. Nevertheless, this initial characterization provides valuable information about deep-sea plasmidome.

The quality of each assembled metagenome and predicted plasmidome was accessed using metaQUAST v5.2.1[65]. Plasmid contig N50 values were subsequently utilized to construct a table expressing length (bp) per sample. Plasmid contig size analysis was performed using stats v4.2.2[66] within the R language v4.2.2[67].

Plasmid profiling

Plasmid sequence contigs were taxonomically classified at the species level using the Platon database. This database integrates data from MPS, RDS, the RefSeq Plasmid database, the PlasmidFinder database, manually curated MOB HMM models from MOBscan, custom conjugation and replication HMM models, and oriT sequences from MOB-suite[68]. The classification was performed using BLAST+ v2.7.1[69]. For taxonomic identification, BLASTn alignments were conducted, and the taxonomy of the plasmid contig was assigned based on the call with the lowest E-value and the highest Total Score.

Subsequently, we mapped the raw reads back to the plasmid contigs using Bowtie2 v2.5.0[57], and Pysam v0.22.0[70] was utilized to obtain read counts for each plasmid sequence contig. The species classification information was then utilized to create feature tables expressing read counts per sample and taxonomy.

Functional annotation and antibiotic resistance genes mining

Plasmid sequence contigs were annotated using Prokka v1.14.6[71] tool. Protein sequences were further analyzed using GhostKOALA[38] to annotate genes and identify metabolic pathways using KEGG reference databases[72]. HTSeq (htseq-count) v2.0.5[73] was employed to quantify the number of mapped reads for each annotated gene in the plasmidome. We constructed a feature table expressing read counts per sample and metabolic pathways, followed by normalization of read counts using the Deseq2 package v1.38.3[74]. The DESeq2 normalization method was chosen because it addresses low dispersion estimates by modeling the relationship between dispersion and average expression across all samples. DESeq2 incorporates several methodological advances and novel features, including the use of shrinkage estimators for dispersion and fold change, which enhance the accuracy and quantitative precision of the analysis[74].

Plasmid sequences contigs were analyzed to predict ARGs using ABRicate v1.0.1[75] with a comprehensive set of databases including ARG-ANNOT[76], CARD[77], EcOH[78], Ecoli_VF[79], MEGARes 2.0[80], NCBI AMRFinderPlus[81], PlasmidFinder[82], ResFinder[83], and VFDB[84]. Subsequently, a feature table was generated expressing read counts per sample for each ARG, utilizing the read counts obtained with Pysam v0.22.0[70]. The Deseq2 v1.38.3[74] package was employed to normalize read counts.

Finally, to get insights about the plasmid mechanisms used in depth-sea environments for the spread of antibiotic resistance and other adaptive traits, plasmid contigs were scanned for MOB family relaxases (enzymes involved in conjugation), using MOB-suite v3.1.7[85]. MOB hits meeting the criteria of E-values of $\leq 1e-5$ and >80% query coverage were included in the analysis. Subsequently, a feature table was generated expressing read counts per sample for each MOB family, utilizing the read counts obtained with Pysam v0.22.0[70]. The Deseq2 v1.38.3[74] package was employed to normalize read counts.

Statistical analysis

Data exploration and statistical analyses were conducted within the R v4.2.2 environment[67] using various R packages including Vegan v2.6[86], stats v4.2.2[66], and Deseq2 v1.38.3[73]. For both whole metagenome and plasmid profiling, sample richness, Shannon entropy, and Pielou's evenness were calculated using Vegan v2.6[86], while Bray-Curtis similarities were computed using the metaMDS() function within the same package. Alpha diversity analysis was performed using Kruskal-Wallis tests with the kruskal.test() function from the stats v4.2.2[66] package, and beta diversity analysis was conducted using PERMANOVA analysis with the adonis() function from Vegan v2.6[86].

To assess the differences in plasmid contig size across depth groups and oceans, Kruskal-Wallis tests were utilized with the kruskal.test() function from the stats v4.2.2[66] package. Normalization of read counts for metabolic pathways, ARGs and MOB families was carried out using the DESeq() function, and data transformation was performed using the normTransform() function from the Deseq2 v1.38.3[73] R package. The DESeq2 package conducts internal normalization by calculating the geometric mean for each gene across all samples and dividing the read counts of each gene by this mean.

All plots were generated using the ggplot2 v3.5.0[87] R package. The complete pipeline (Fig. S13), utilized Python scripts for constructing feature tables, while R scripts were employed for further analysis. Both Python and R scripts are available online on GitLab: https://gitlab.com/CNCA_CeNAT/deep-sea-plasmidomes

Data availability

The data and pipeline needed to reproduce this work are available at GitLab: https://gitlab.com/CNCA_CeNAT/deep-sea-plasmidomes

Received: 1 September 2024; Accepted: 8 January 2025
Published online: 23 January 2025

References

1. Arnold, B., Huang, I. & Hanage, W. Horizontal gene transfer and adaptive evolution in bacteria. *Nat. Rev. Microbiol.* **20**, 206–218. https://doi.org/10.1038/s41579-021-00650-4 (2022).
2. Rankin, D., Rocha, E. & Brown, S. What traits are carried on mobile genetic elements, and why?. *Heredity*[SPACE]https://doi.org/10.1038/hdy.2010.24 (2010).
3. Androsiuk, L., Shay, T. & Tal, S. Characterization of the environmental plasmidome of the red sea. *Microbiol. Spectrum*[SPACE]https://doi.org/10.1128/spectrum.00400-23 (2023).
4. Garcillán-Barcia, M., Redondo-Salvo, S. & de la Cruz, F. Plasmid classifications. *Plasmid* **126**, 102684. https://doi.org/10.1016/j.plasmid.2023.102684 (2023).
5. Finks, S. & Martiny, J. Plasmid-encoded traits vary across environments. *mBio*[SPACE]https://doi.org/10.1128/mbio.03191-22 (2023).
6. Brockhurst, M. & Harrison, E. Ecological and evolutionary solutions to the plasmid paradox. *Trends Microbiol.* **30**, 534. https://doi.org/10.1016/j.tim.2021.11.001 (2022).
7. Lerminiaux, N. & Cameron, A. Horizontal transfer of antibiotic resistance genes in clinical environments. *Can. J. Microbiol.* **65**, 34. https://doi.org/10.1139/cjm-2018-0275 (2018).
8. Smalla, K., Jechalke, S. & Top, E. Plasmid detection, characterization, and ecology. *Microbiol. Spectrum*[SPACE]https://doi.org/10.1128/microbiolspec.plas-0038-2014 (2015).
9. Kerkhof, L. & Goodman, R. Ocean microbial metagenomics. *Deep Sea Res. Part II* **56**, 1824. https://doi.org/10.1016/j.dsr2.2009.05.005 (2009).
10. Blau, K., Jechalke, S. & Smalla, K. Detection, isolation, and characterization of plasmids in the environment. *Methods Mol. Biol.* **2075**, 39–60. https://doi.org/10.1007/978-1-4939-9877-7_3 (2020).
11. Sobecky, P. A. Approaches to investigating the ecology of plasmids in marine bacterial communities. *Plasmid* **48**, 213–221. https://doi.org/10.1016/s0147-619x(02)00110-5 (2002).
12. Orlek, A. et al. Plasmid classification in an era of whole-genome sequencing: Application in studies of antibiotic resistance epidemiology. *Front. Microbiol.*[SPACE]https://doi.org/10.3389/fmicb.2017.00182 (2017).
13. Perez, M. F. et al. First report on the plasmidome from a high-altitude lake of the andean puna. *Front. Microbiol.*[SPACE]https://doi.org/10.3389/fmicb.2020.01343 (2020).
14. Cuadrat, R. R. C., Sorokina, M., Andrade, B. G., Goris, T. & Dávila, A. M. R. Global ocean resistome revealed: Exploring antibiotic resistance gene abundance and distribution in tara oceans samples. *GigaScience*[SPACE]https://doi.org/10.1093/gigascience/giaa046 (2024).
15. Rogers, A. D. Environmental change in the deep ocean. *Annu. Rev. Environ. Resour.* **40**, 1–38. https://doi.org/10.1146/annurev-environ-102014-021406 (2015).
16. Sanganyado, E., Chingono, K. E., Gwenzi, W., Chaukura, N. & Liu, W. Organic pollutants in deep sea: Occurrence, fate, and ecological implications. *Water Res.*[SPACE]https://doi.org/10.1016/j.watres.2021.117658 (2021).
17. Abel, S., Wu, F., Primpke, S., Gerdts, G. & Brandt, A. Journey with no return: Plastic pollution in the hadal of deep-sea trenches. *SSRN Electron. J.*[SPACE]https://doi.org/10.2139/ssrn.4180311 (2022).
18. Zhu, L. et al. Microplastic ingestion in deep-sea fish from the south China sea. *Sci. Total Environ.* **677**, 493–501. https://doi.org/10.1016/j.scitotenv.2019.04.380 (2019).
19. Gilbert, N. Complex deep-sea expeditions try to size up seabed mining impacts. *Proc. Natl. Acad. Sci. USA*[SPACE]https://doi.org/10.1073/pnas.2404667121 (2024).
20. Levin, L. A., Amon, D. J. & Lily, H. Challenges to the sustainability of deep-seabed mining. *Nat. Sustain.* **3**, 784. https://doi.org/10.1038/s41893-020-0558-x (2020).
21. Thompson, K. F., Miller, K. A., Currie, D., Johnston, P. & Santillo, D. Seabed mining and approaches to governance of the deep seabed. *Front. Mar. Sci.*[SPACE]https://doi.org/10.3389/fmars.2018.00480 (2018).
22. Dib, J. R., Wagenknecht, M., Farías, M. E. & Meinhardt, F. Strategies and approaches in plasmidome studies-uncovering plasmid diversity disregarding of linear elements?. *Front. Microbiol.*[SPACE]https://doi.org/10.3389/fmicb.2015.00463 (2015).
23. Gao, X., Fu, X., Xie, M. & Wang, L. Environmental risks of antibiotic resistance genes released from biological laboratories and its control measure. *Environ. Monit. Assess.*[SPACE]https://doi.org/10.1007/s10661-023-11316-4 (2023).
24. Acinas, S. G. et al. Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. *Commun. Biol.* **4**, 604. https://doi.org/10.1038/s42003-021-02112-2 (2021).
25. Pesant, S. et al. Open science resources for the discovery and analysis of tara oceans data. *Sci. Data*[SPACE]https://doi.org/10.1038/sdata.2015.23 (2015).
26. Rigonato, J. et al. Ocean-wide comparisons of mesopelagic planktonic community structures. *ISME Commun.*[SPACE]https://doi.org/10.1038/s43705-023-00217-y (2023).
27. Salazar, G. et al. Global diversity and biogeography of deep-sea pelagic prokaryotes. *ISME J.* **10**, 596. https://doi.org/10.1038/ismej.2015.137 (2024).
28. Sanz-Sáez, I. et al. Microorganisms involved in methylmercury demethylation and mercury reduction are widely distributed and active in the bathypelagic deep ocean waters. *Environ. Sci. Technol.* **58**, 13795. https://doi.org/10.1021/acs.est.3c05743 (2024).
29. Sunagawa, S. et al. Structure and function of the global ocean microbiome. *Science* **348**, 1261359. https://doi.org/10.1126/science.1261359 (2015).
30. Brown Kav, A. et al. Unravelling plasmidome distribution and interaction with its hosting microbiome. *Environ. Microbiol.* **22**, 32. https://doi.org/10.1111/1462-2920.14813 (2019).
31. Schwengers, O. et al. Platon: Identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microb. Genom.*[SPACE]https://doi.org/10.1099/mgen.0.000398 (2020).
32. Yu, Z. et al. Insights into the circular: The cryptic plasmidome and its derived antibiotic resistome in the urban water systems. *Environ. Int.*[SPACE]https://doi.org/10.1016/j.envint.2023.108351 (2024).
33. Ma, Y., Paulsen, I. T. & Palenik, B. Analysis of two marine metagenomes reveals the diversity of plasmids in oceanic environments. *Environ. Microbiol.* **14**, 453. https://doi.org/10.1111/j.1462-2920.2011.02633.x (2011).
34. Sobecky, P. A. Approaches to investigating the ecology of plasmids in marine bacterial communities. *Appl. Environ. Microbiol.* **63**, 888–895. https://doi.org/10.1128/aem.63.3.888-895.1997 (1997).
35. Goodarzi, Z., Asad, S. & Mehrshad, M. Genome-resolved insight into the reservoir of antibiotic resistance genes in aquatic microbial community. *Sci. Rep.*[SPACE]https://doi.org/10.1038/s41598-022-25026-3 (2022).

36. Gilbert, J. A. et al. Defining seasonal marine microbial community dynamics. *ISME J.* **6**, 298. https://doi.org/10.1038/ismej.2011.107 (2011).
37. Grzymski, J. J. et al. Comparative genomics of dna fragments from six antarctic marine planktonic bacteria. *Appl. Environ. Microbiol.* **72**, 1532–1541. https://doi.org/10.1128/aem.72.2.1532-1541.2006 (2006).
38. Kanehisa, M., Sato, Y. & Morishima, K. Blastkoala and ghostkoala: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* **428**, 726–731. https://doi.org/10.1016/j.jmb.2015.11.006 (2016).
39. Perez, M. et al. First report on the plasmidome from a high-altitude lake of the andean puna. *Front Microbiol.*[SPACE]https://doi.org/10.3389/fmicb.2020.01343 (2020).
40. Ankita, K., Wu, Y.-W., & Chandonia, J.-M. et al. Large circular plasmids from groundwater plasmidomes span multiple incompatibility groups and are enriched in multimetal resistance genes. *mBio*. https://doi.org/10.1128/mbio.02899-18 (2019).
41. Sentchilo, V. et al. Community-wide plasmid gene mobilization and selection. *ISME J.* **7**, 1173–1186. https://doi.org/10.1038/ismej.2013.13 (2013).
42. Putman, L. et al. Deconstructed plastic substrate preferences of microbial populations from the natural environment. *Microbiol. Spectr.*[SPACE]https://doi.org/10.1128/spectrum.00362-23 (2023).
43. Gutierrez, T. Marine, aerobic hydrocarbon-degrading gammaproteobacteria: Overview. In *Taxonomy, Genomics and Ecophysiology of Hydrocarbon-Degrading Microbes*, 1. https://doi.org/10.1007/978-3-319-60053-6_22-1 (Handbook of Hydrocarbon and Lipid Microbiology, 2017).
44. Georgieva, M., Mevenkamp, L. & Paulus, E. Shedding light on deep-sea biodiversity-a highly vulnerable habitat in the face of anthropogenic change. *Front. Mar. Sci.*[SPACE]https://doi.org/10.3389/fmars.2021.667048 (2021).
45. Sackeyfio, N. Spatio-temporal analysis of global marine oil tanker spills over half a century. https://www.itopf.org/knowledge-resources/documents-guides/spatio-temporal-analysis-of-global-marine-oil-tanker-spills-over-half-a-century-2020/ (2020). [Accessed 11 October 2022].
46. Pabortsava, K. & Lampitt, R. S. High concentrations of plastic hidden beneath the surface of the Atlantic Ocean. *Nat. Commun.* **11**, 4411. https://doi.org/10.1038/s41467-020-18206-6 (2020).
47. Zhao, C. et al. Monitoring and evaluation of antibiotic resistance genes in three rivers in Northeast China. *Environ. Sci. Pollut. Res.* **29**, 44148. https://doi.org/10.1007/s11356-022-18555-x (2022).
48. Dzelalija, M. et al. Large-scale biogeographical shifts of abundance of antibiotic resistance genes and marine bacterial communities as their carriers along a trophic gradient. *IJMS*[SPACE]https://doi.org/10.3390/ijms25010654 (2024).
49. Zhang, H. et al. Unveiling the occurrence, hosts and mobility potential of antibiotic resistance genes in the deep ocean. *Sci. Total Environ.* **816**, 151539. https://doi.org/10.1016/j.scitotenv.2021.151539 (2022).
50. Smillie, C., Garcillán-Barcia, M. P., Francia, M. V., Rocha, E. P. & De La Cruz, F. Mobility of plasmids. *Microbiol. Mol. Biol. Rev.* **74**, 434. https://doi.org/10.1128/mmbr.00020-10 (2010).
51. Benoit, G. et al. SimkaMin: Fast and resource frugal de novo comparative metagenomics. *Bioinformatics* **36**, 1275–1276. https://doi.org/10.1093/bioinformatics/btz685 (2020).
52. Benoit, G. et al. Multiple comparative metagenomics using multiset k-mer counting. *PeerJ Comput. Sci.* **2**, e94. https://doi.org/10.7717/peerj-cs.94 (2016).
53. Leinonen, R. et al. The European nucleotide archive. *Nucleic Acids Res.* **39**, D28. https://doi.org/10.1093/nar/gkq967 (2010).
54. National Center for Biotechnology Information (NCBI). National Center for Biotechnology Information (NCBI). (1988). Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information.
55. Babraham Bioinformatics. Fastqc: A quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed: 2024-06-21.
56. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one fastq preprocessor. *Bioinformatics* **34**, i884–i890. https://doi.org/10.1093/bioinformatics/bty560 (2018).
57. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25. https://doi.org/10.1186/gb-2009-10-3-r25 (2009).
58. Genome taxonomy database (gtdb). https://gtdb.ecogenomic.org/.
59. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with kraken 2. *Genome Biol.* **20**, 257. https://doi.org/10.1186/s13059-019-1891-0 (2019).
60. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: Estimating species abundance in metagenomics data. *PeerJ Comput. Sci.*[SPACE]https://doi.org/10.7717/peerj-cs.104 (2017).
61. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaspades: A new versatile metagenomic assembler. *Genome Res.* **27**, 824. https://doi.org/10.1101/gr.213959.116 (2017).
62. Bushnell, B. BBMap: A fast, accurate, splice-aware aligner. (2014). Available from: https://sourceforge.net/projects/bbmap/.
63. Rojas-Villalta, D., Calderón-Osorno, M., Barrantes-Jiménez, K., Arias-Andres, M. & Rojas-Jiménez, K. Benchmarking ai-based plasmid annotation tools for antibiotic resistance genes mining from metagenome of the virilla river, costa rica. In *2023 IEEE 5th International Conference on BioInspired Processing (BIP)*, 1–6. https://doi.org/10.1109/BIP60195.2023.10379392 (2023).
64. Nishimura, Y. et al. Environmental viral genomes shed new light on virus-host interactions in the ocean. *MSphere* **2**, e00359-16. https://doi.org/10.1128/msphere.00359-16 (2017).
65. Mikheenko, A., Saveliev, V. & Gurevich, A. Metaquast: Evaluation of metagenome assemblies. *Bioinformatics* **32**, 1088–1090. https://doi.org/10.1093/bioinformatics/btv697 (2016).
66. R Core Team. *Stats package*. R Foundation for Statistical Computing (2019). R package Version 4.2.2.
67. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2023).
68. Schwengers, O. Platon database. (2020). Available from: https://zenodo.org/records/3751774.
69. Camacho, C. et al. Blast+: Architecture and applications. *BMC Bioinf.*[SPACE]https://doi.org/10.1186/1471-2105-10-421 (2009).
70. Heger, A. et al. pysam: htslib interface for python. (2023). Available from: https://pysam.readthedocs.io/en/latest/api.html.
71. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069. https://doi.org/10.1093/bioinformatics/btu153 (2014).
72. Kanehisa, M. & Goto, S. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30. https://doi.org/10.1093/nar/28.1.27 (2000).
73. Anders, S., Pyl, P. T. & Huber, W. Htseq—A python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166. https://doi.org/10.1093/bioinformatics/btu638 (2014).
74. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. *Genome Biol.*[SPACE]https://doi.org/10.1186/s13059-014-0550-8 (2014).
75. Seemann, T. ABRicate. Available from: https://github.com/tseemann/abricate.
76. Gupta, S. K. et al. Arg-annot, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob. Agents Chemother.* **58**, 212–220. https://doi.org/10.1128/aac.01310-13 (2014).
77. Jia, B. et al. Card 2017: Expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **45**, D566–D573. https://doi.org/10.1093/nar/gkw1004 (2017).
78. Ingle, D. J., Valcanis, M., Kuzevski, A. et al. In silico serotyping of *E. coli* from short read data identifies limited novel o-loci but extensive diversity of o:h serotype combinations within and between pathogenic lineages. *Microb. Genom.* https://doi.org/10.1099/mgen.0.000064 (2016).

79. National Microbiology Laboratory. Escherichia coli virulence factors. Available from: https://github.com/phac-nml/ecoli_vf.
80. Doster, E. et al. Megares 2.0: A database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. *Nucleic Acids Res.* **48**, D561–D569. https://doi.org/10.1093/nar/gkz1010 (2020).
81. Feldgarden, M. et al. Validating the amrfinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrob. Agents Chemother.*[SPACE]https://doi.org/10.1128/aac.00483-19 (2019).
82. Alessandra, C. et al. In silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* **58**, 3895–3903. https://doi.org/10.1128/aac.02412-14 (2014).
83. Zankari, E. et al. Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* **67**, 2640–2644. https://doi.org/10.1093/jac/dks261 (2012).
84. Chen, L., Zheng, D., Liu, B., Yang, J. & Jin, Q. Vfdb 2016: Hierarchical and refined dataset for big data analysis-10 years on. *Nucleic Acids Res.* **44**, D694–D697. https://doi.org/10.1093/nar/gkv1239 (2016).
85. Robertson, J. & Nash, J. H. E. MOB-suite: Software tools for clustering, reconstruction and typing of plasmids from draft assemblies. https://doi.org/10.6084/m9.figshare.6177188 (2018).
86. Oksanen, F. J. and et al. *Vegan: Community Ecology Package. R package Version 2.6-4* (2017).
87. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, New York, 2009).

Acknowledgements

This research was partially supported by computational resources provided by the Kabré supercomputer at the Costa Rica National High Technology Center. K.R.J. was supported by the Vicerrectoría de Investigacion de la Universidad Costa Rica (projects C1455, C2650, C3509). M.A.A. was supported by the Universidad Nacional, Costa Rica (Project SIA 0483-21) and The Research Council of Norway, Marmib Project 315812.

Author contributions

M.C.O, and K.R.J. designed research; M.C.O, D.R.V, and K.R.J. performed research; M.C.O, D.R.V, F.L. and K.R.J. analyzed data; and M.C.O, D.R.V, F.L., J.C., M.A.A., and K.R.J wrote the paper. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-86098-5.

Correspondence and requests for materials should be addressed to M.C.-O. or K.R.-J.