



OPEN

Metagenomic exploration of Andaman region of the Indian Ocean

Vishnu Prasoodanan P. K.^{1,2}, Sudhir Kumar^{1,2}, Darshan B. Dhakan¹, Prashant Waiker¹, Rituja Saxena¹ & Vineet K. Sharma¹✉

Ocean microbiome is crucial for global biogeochemical cycles and primary productivity. Despite numerous studies investigating the global ocean microbiomes, the microbiome composition of the Andaman region of the Indian Ocean remains largely unexplored. While this region harbors pristine biological diversity, the escalating anthropogenic activities along coastal habitats exert an influence on the microbial ecology and impact the aquatic ecosystems. We investigated the microbiome composition in the coastal waters of the Andaman Islands by 16S rRNA gene amplicon and metagenomic shotgun sequencing approaches and compared it with the Tara Oceans Consortium. In the coastal waters of the Andaman Islands, a significantly higher abundance and diversity of *Synechococcus* species was observed with a higher abundance of photosynthesis pigment-related genes to adapt to variable light conditions and nutrition. In contrast, *Prochlorococcus* species showed higher abundance in open ocean water samples of the Indian Ocean region, with a relatively limited functional diversity. A higher abundance of antibiotic-resistance genes was also noted in the coastal waters region. We also updated the ocean microbiome gene catalog with 93,172 unique genes from the Andaman coastal water microbiome. This study provides valuable insights into the Indian Ocean microbiome and supplements the global marine microbial ecosystem studies.

Oceans harbor diverse life forms that play a crucial role in global biogeochemical fluxes including nitrogen and carbon cycling. Phytoplankton, such as single-celled algae and cyanobacteria, are prevalent in surface waters and contribute significantly to carbon fixation¹. Similarly, nitrogen input in oceanic ecosystems is primarily facilitated by bacteria such as *Trichodesmium* and *Crocosphaera* (diazotrophs) that convert atmospheric nitrogen into ammonium which is essential for making amino acids and proteins in all forms of life². Oceanic microbes also influence the cycling of essential nutrients like iron, phosphorus, and sulfur^{3,4}.

Cyanobacteria, specifically the dominant genera like *Synechococcus* and *Prochlorococcus*, are the primary autotrophic prokaryotes in marine surface waters that enrich the photosynthetic biomass and play a significant role in ecosystem functioning⁵. Even though both are ubiquitous, their abundance varies on local scales and both have adaptations to the local environmental conditions^{6–8}. *Synechococcus* species have evolved and accumulated a wide variety of pigmentation that allow them to grow in varying light conditions in well-mixed coastal environments⁹. In contrast, *Prochlorococcus* is perhaps the most abundant photosynthetic organism on earth and can adapt to low nutrient conditions in oligotrophic oceans depleted of macronutrients⁸.

In recent years, next-generation sequencing and integration of multi-omics approaches have advanced the understanding of taxonomic and functional composition of ocean microbiome. Craig Venter pioneered the use of metagenomic sequencing for studying ocean microbial diversity. They employed shotgun sequencing techniques to obtain about 1 Gbp of sequencing data from the Sargasso Sea near Bermuda¹⁰. Subsequently, the Sorcerer II Global Ocean Sampling (GOS) project conducted sampling along the North American coast, eastern tropical Pacific, and equatorial Indian Ocean, focusing on surface waters¹¹. Later on, Beatriz Díez et al. studied the diversity and functional aspects of picocyanobacteria in open waters of the equatorial Indian Ocean¹². Further studies in the field explored the deep-water samples to understand the effects of factors such as sunlight, pressure, and oxygen levels that vary with depth and their influence on microbial diversity^{13–16}.

A large-scale effort for a holistic understanding of marine water communities was conducted through the Tara Oceans expedition, which collected samples from diverse global locations and generated extensive metagenomic sequence data (<https://ocean-microbiome.embl.de/companion.html>). This expedition covered the Pacific

¹MetaBioSys Group, Department of Biological Sciences, Indian Institute of Science Education and Research Bhopal, Bhopal, India. ²These authors contributed equally: Vishnu Prasoodanan P. K. and Sudhir Kumar. ✉email: vineetks@iiserb.ac.in

Ocean, the North and South Atlantic Oceans, and the western parts of the Indian Ocean, specifically the Arabian Sea within the Indian Monsoon Gyres Province (MONS) and Northwestern Arabian Sea Upwelling Province (ARAB). The data obtained from this expedition contributed to the construction of a reference metagenomic gene catalog from marine water environments¹⁷.

To date, the microbiome composition of the eastern part of the Indian Ocean remains largely unexplored. The studies conducted in the Indian Ocean region were either limited to amplicon-based sequencing, which lacks comprehensive functional representation of microbial communities, or their sampling was confined to the equatorial Indian Ocean region^{12,14–16}. In addition, the majority of previous studies were focused on the microbiome of open ocean regions. There are a few reports of culture-based exploration in these areas¹⁸, however, the microbial composition of coastal waters remains largely understudied.

The Andaman and Nicobar Islands region in the Indian Ocean has diverse terrestrial and marine ecosystems like mangroves, coral reefs, beaches, etc. The increasing impact of various human activities in the coastal areas, which primarily contribute to the aquatic food-animal supply, could be a potential source of antibiotic resistance transmission^{19,20}. The majority of antibiotics are excreted into the environment and end up in aquatic ecosystems, which serve as a reservoir of antibiotic resistance genes (ARGs)^{21–23}. Thus, it is important to explore the microbiome composition and genetic makeup of coastal marine ecological units. This diversity also holds substantial potential for biomedical applications^{24,25}.

We carried out a comprehensive study to understand the diversity and functional roles of microbial communities in coastal waters from the Andaman Islands which is a yet unexplored part of the Indian Ocean region. We employed a combined approach of 16S rRNA amplicon and whole metagenomic shotgun sequencing to explore microbiome composition and its functional potential. We also used previously available metagenomic data from the Tara Oceans study for comparative analysis of coastal with open ocean microbiome. This study provides the first compositional and functional characterization of the microbial communities in the coastal water region of the Andaman Sea.

Methods

Andaman data description

Marine water samples were collected from Andaman and Nicobar Islands, India located at 11.61 °N, 92.74 °E, and Cape-Comorin, the southernmost point of peninsular India. Water samples ranging from 2 to 3 L were collected using sterile plastic cans/carboys and transported to the laboratory in ice packs. All marine water samples were subjected to filtration with a 1.2 µm pore size membrane to avoid contamination of other micro-eukaryotes and suspended organic matter²⁶. Nine samples from six locations at variable distances from the coast of the Andaman islands and one sample from Cape-Comorin were collected for 16S rRNA amplicon sequencing. Additionally, the same six samples from each location in the Andaman region were used for whole metagenome sequencing.

DNA extraction and sequencing

Coarse particles and debris were removed from the samples by filtering through a 1.2 µm membrane filter. To entrap the prokaryotic cells the filtrate was passed through a 0.22 µm pore-size membrane filter²⁷. DNA was extracted from this membrane using the metagenomic DNA isolation kit for water (Epicenter, Wisconsin, USA), according to the manufacturer's protocol with minor modifications like the addition of 100 µl of 5 M NaCl to 700 µl isopropanol for efficient precipitation of DNA. The DNA pellet was resuspended in 10 mM Tris (pH 8.5) and evaluated on a Geneva nanodrop micro-spectrophotometer (Jenway, Bibby Scientific Limited, U.K) and Qubit HS dsDNA kit (Life technologies, USA). We generated an amplicon library targeting the V3 region of the 16S rRNA gene, using purified DNA as a template. This specific region, along with V4 and V3-V4, has demonstrated high accuracy in discerning taxonomic diversity within bacteria^{28,29}. The details of primers used for amplification are shared in supplementary data (Supplementary Table 3). PCR was performed with initial denaturation at 94 °C for 5 min, followed by 35 cycles of denaturation at 94 °C for 30 s, annealing at 69 °C for 30 s, extension at 72 °C for 30 s, and final extension cycle for 5 min at 72 °C. Taq DNA Polymerase (Life technologies, USA) was used and 5% DMSO was added to the master mix to enhance the concentration of amplified product from the metagenomic GC-rich template. The amplified products were evaluated on 2% w/v agarose gel and purified using the Agencourt Ampure XP kit (Beckman Coulter, USA). The libraries were prepared using the Illumina 16S metagenomic library preparation guide. The size of prepared library was estimated on 2100 Bioanalyzer using Bioanalyzer DNA 1000 kit (Agilent, USA) and library concentration was measured with Qubit 2.0 fluorometer using dsDNA HS kit (Life technologies, USA).

Shotgun metagenomics libraries were prepared using the Illumina Nextera XT sample preparation kit (Illumina, USA) following the manufacturer's protocol. The libraries were evaluated on 2100 Bioanalyzer using Bioanalyzer DNA 1000 kit (Agilent, USA) to estimate the library size. Library was quantified using Qubit dsDNA HS kit (Life technologies, USA) and KAPA SYBR Fast qPCR Master mix, Illumina standards, and primer premix (kappa Biosystems, USA). Paired-end sequencing (150 bp) of both libraries was performed on the Illumina NextSeq500 platform (Illumina, USA).

Collection of publicly available open ocean microbiome data from Tara Oceans Consortium

Open ocean samples from Tara Oceans Consortium¹⁷ were retrieved for the comparative analysis of community structure with the coastal microbiome. Three criteria were employed for selecting samples from the Tara Ocean Consortium data, as described in the following text. (i) Sample collected from Marine pelagic biomes of ARAB and MONS were used for the analysis, (ii) Only surface water samples were considered (sampling depth = 5 m) to allow relevant comparisons with our data, (iii) Only samples from size fractions of 0.1–0.22 µm and 0.22–1.6 µm were taken into consideration and size fractions of <0.22 µm were excluded. Using these criteria, a total of

985,580,126 reads ($109,508,902.88 \pm 83,144,900.59$, mean \pm sd) from nine selected samples (from six stations) of publicly available Tara Oceans Consortium data were used for the analysis.

16S rRNA amplicon analysis

A total of 17,604,113 ($1,956,012.56 \pm 645,294.51$, mean \pm sd) paired-end reads from nine water samples of Andaman region and 172,074,584 paired-end reads from Cape-Comorin water sample were generated. The reads were filtered to remove ambiguous bases using AmbiguityFiltering.pl of the NGS-QC toolkit³⁰, no ambiguous bases were allowed, and reads shorter than 60 bp were discarded (-c 0, -n 60). Reads with 70% bases above Q20 were filtered using IlluQC_PRL.pl. Primers were removed using cutadapt 1.18³¹, and remaining single-end reads post-filtration were eliminated using the repair.sh function of the bmap toolkit, later the paired reads were assembled using FLASH (Fast Length Adjustment of SHort reads)³². Merged paired-end reads were imported in QIIME-2 for further analysis³³. The reads were denoised and filtered for chimeric reads using dada2³⁴. Low abundant Amplicon Sequence Variants (ASVs/features) with a total frequency of 10 across samples were removed and the resulting feature table was used for further analysis. Taxonomic annotation of ASVs was carried out using the q2-feature-classifier plugin after training on V3 region using Greengenes 13_5 reference database³⁵. After taxonomic assignment, the relative abundance of phylum, genus, and species in each sample was calculated. Observed features and Shannon diversity were calculated to estimate alpha diversity. Inter-sample distances were calculated using Jaccard, Bray–Curtis, and UniFrac distances.

Metagenomic data analysis

A total of 73,165,072 ($12,194,178.67 \pm 3,638,208.02$) paired-end reads were generated from six samples from Andaman region. In addition, 985,580,126 paired-end reads ($109,508,903 \pm 83,144,900.59$) were retrieved from nine selected samples of Tara Oceans Consortium. Metagenomic reads were filtered to remove ambiguous bases using AmbiguityFiltering.pl of the NGS-QC toolkit allowing zero ambiguous bases and sequences shorter than 60 bases were discarded (-c 0 -n 60). The resulting reads were filtered to remove low-quality reads. Further, quality filtration of reads was carried out with a cut-off of $\geq Q25$ (N 1 -l 70 -s 25). Assembly of metagenomic reads from each sample was carried out using SOAPdenovo³⁶ at different k-mers (53 to 109) and selected the best-assembled contigs for each sample based on different statistical parameters (Supplementary Fig. 1).

Taxonomic assignment of metagenomic reads

Taxonomic annotation of metagenomic reads was carried out using Kaiju and Kraken2^{37,38}. Kaiju classified 692,372,913 of 985,093,743 (70.28%) reads using default parameters. Kraken2 classified 340,218,793 of 985,093,743 (34.54%) reads. Species-level sequence abundance was inferred using Bracken³⁹. Microbial genera present in all samples of coastal and open ocean groups with > 1% relative abundance were considered for defining core-microbiome composition.

Reconstruction of ocean microbiome gene catalog

Open reading frames (ORFs) were predicted from contigs (contig size > 300 bp) of each sample separately using MetaGeneMark version 3.38⁴⁰. The Ocean Microbial Reference Gene Catalog (OMRGC) consisting of 40,154,822 genes (from 243 ocean samples) constructed by the Tara Oceans Consortium was retrieved. For assessing the microbial gene diversity of the Indian Ocean region, we added 113,410 ORFs (> 100 bp nucleotide) predicted from Andaman data to OMRGC and carried out clustering of redundant set of genes using CD-HIT at 100% amino acid sequence identity and selected the longest gene from each cluster as a representative in updated non-redundant gene catalog⁴¹. The final non-redundant gene catalog consisted of 39,732,487 genes. Since differences in sequencing depth can create biases in gene quantification and comparative analysis, we subsampled 10 million reads (estimated based on Andaman data rarefaction) from Tara Oceans Consortium data and used those for further analysis.

Gene quantification in Andaman and Tara Oceans samples and functional analysis

High-quality reads from Andaman and Tara Oceans Consortium samples (after subsampling) were mapped to the updated ocean microbiome gene catalog using soap2 of SOAPaligner version 2.21 with an identity cutoff of $\geq 90\%$ ⁴². Alignments of two types were considered for sequence-based profiling: (i) Entire paired-end read mapped to gene, and (ii) One end of paired-end reads mapped and the other read remained unmapped. In both cases, the read was counted as one copy. Further, the read count was normalized based on the length of the gene as $b_i = x_i/L_i$. The relative abundance of a gene within the sample was calculated as follows:

$$a_i = b_i / \sum_j b_j = x_i/L_i / \sum_j x_j/L_j$$

where a_i is the relative abundance of gene in sample S , x_i is the number of time gene i was detected in sample S (i.e. the number of mapped reads), L_i is the length of gene i , j is all the genes (j number of genes) and b_i is copy number of gene i in sequence data from sample S ^{43,44}.

A total of 5,493,722 genes were present in at least one out of 15 samples. We filtered the quantified genes based on the two criteria: (i) genes should be present in more than 20% of the samples (3 out of 15 samples), and (ii) genes with > 0.05% relative abundance was reincorporated to avoid losing highly abundant genes that might be excluded in the previous criteria. After applying these gene exclusion criteria, 1,867,938 genes were selected and used for further analysis. Functional annotation of metagenomic reads and genes is needed to gain functional insights^{45–47} and was performed for 5,493,722 genes using eggNOG-mapper v2 with eggNOG

5.0 database using diamond at default parameters^{48,49}. The cumulative abundance of each KO and CAZy gene family was also calculated.

Identification of antibiotic resistance genes

Antibiotic resistance genes were identified using Resistance Gene Identifier (RGI)⁵⁰. From the three different prediction criteria provided by RGI i.e., perfect, strict, and loose options based on different types of hits in homology search, we used the “strict” criteria for the detection of ARGs in the updated ocean microbiome gene catalog. We also calculated the cumulative abundance of each Antibiotic Resistance Ontology (ARO), AMR gene families, drug classes, and antibiotic resistance mechanisms of ARGs detected in coastal water and open ocean samples.

Analyses based on *Synechococcus* and *Prochlorococcus* genomes

We retrieved 1100 *Prochlorococcus* genomes and 405 *Synechococcus* genomes from the NCBI GenBank database (including complete and draft genomes). Custom databases were constructed for both genera separately using the above-mentioned genomes. The databases were indexed using bowtie2 and the abundance of each genome in samples was quantified by mapping high-quality metagenomic reads from Andaman and Tara Oceans Consortium samples against the custom genome database using bowtie-2. We proceeded with analyzing the top 100 abundant genomes from both coastal and open ocean samples. *Synechococcus* represented 85 out of 100 top abundant genomes in coastal samples, while 81 out of 100 top abundant genomes in open ocean samples were from the *Prochlorococcus* genus. Open reading frames were predicted from all 166 genomes (81 + 85) using Prodigal. We constructed a non-redundant gene set for both *Prochlorococcus* and *Synechococcus* separately by clustering the genes using CD-HIT at 100% amino acid identity^{41,51}. Functional annotation of both gene sets was carried out using eggNOG-mapper v2⁴⁸. We also examined the orthologous genes present in both gene sets using OrthoFinder and OrthoVenn2^{52,53}. We also predicted 16S rRNA gene sequences from these highly abundant *Synechococcus* (n = 85) and *Prochlorococcus* (n = 81) genomes using barnap (<https://github.com/tseem/ann/barnap>) to evaluate the phylogenetic relatedness/structure of highly abundant genomes in coastal and open ocean waters⁵⁴. Only 45 and 70 16S rRNA gene sequences could be predicted from genomes of *Synechococcus* and *Prochlorococcus* respectively. The complete 16S rRNA gene sequences were aligned using the Clustal-W algorithm implemented in MEGA-X⁵⁵. We also constructed a phylogenetic tree using the neighbor-joining method by introducing *E. coli* 16S rRNA gene as an outgroup.

Statistical analysis

Alpha and beta diversity of coastal and open oceans samples based on the taxonomic and functional composition of microbiomes by 16S rRNA amplicon and metagenome data were carried out using QIIME-2. Principal Coordinate Analysis (PCoA) was carried out using QIIME-2³³. Differentially abundant species/genera were identified using LEfSe and Boruta^{56,57}. The maps used in Fig. 1 were generated using ArcGIS Online (Esri). Plots were made using ggplot2 package in R⁵⁸.

Results

We collected surface water samples from six different coastal locations of the Andaman and Nicobar Islands and employed whole metagenomic and 16S rRNA gene amplicon-based approaches to study the microbiome composition of coastal water. Nine open ocean surface water samples of the Indian Ocean region from the Tara Oceans Consortium were selected for the comparative analysis (Supplementary Table 1, see methods for detailed information on sample collection). We generated whole metagenomic data from six samples of Andaman comprising 73,165,072 (12,194,178.67 ± 3,638,208.02) paired-end reads. A total of 985,580,126 (Average 109,508,903 ± 83,144,900.59) paired-end reads were retrieved for surface samples of the Indian Ocean region from Tara Oceans Consortium for comparative analysis (Supplementary Tables 1, 2, Fig. 1A, See Methods section).

We carried out 16S rRNA sequencing and analysis of ten samples including nine samples collected from coastal regions in Andaman and one sample from Cape-Comorin (CC). A total of 17,604,113 (1,956,012.56 ± 645,294.51) and 172,074,584 paired-end reads were generated for Andaman and CC respectively. Since the sequencing data from CC has significantly higher coverage, we extracted 5,562,311 reads (randomly) from the sample as a representative to avoid any bias in the analysis (Supplementary Table 3, Fig. 1A). After denoising and chimera removal using QIIME-2, 17,105,642 reads (90.77%) remained for further analysis. The analysis identified 23,303 ASVs present across all ten samples. Low-abundant ASVs were excluded and the remaining 19,604 features (ASVs) were used for calculating diversity matrices and examining taxonomic composition (See “Methods” section).

Microbial richness and diversity of coastal water samples

Diversity and richness of Andaman and CC samples were evaluated using Shannon evenness, number of observed features (richness), Pielou’s evenness, and phylogenetic diversity (Faith’s PD) index. Higher microbial diversity (using Shannon evenness, number of observed features, and Pielou’s evenness) was observed at LOC3. The phylogenetic diversity (Faith’s PD) index showed the highest alpha diversity at CC. A variation of alpha diversity was observed in different sampling depths of LOC2 sites with LOC2-4 having the highest diversity (Supplementary Table 4, Supplementary Fig. 2A).

PCoA based on inter-sample distances (Bray–Curtis, Jaccard, and UniFrac) derived from 16S data showed close clustering of LOC2-1, LOC2-2, and LOC2-3, reasonably due to their geographical proximity. Andaman samples clustered far apart from the CC sample (Fig. 1B). PCoA based on inter-sample distances calculated using species abundance from metagenomic data resulted in distinct clustering of coastal and open ocean samples (Fig. 1C). The first and second principal coordinates explained 57.07 and 20.61% variations, respectively. These observations indicate the impact of geographical locations on overall microbiome composition.

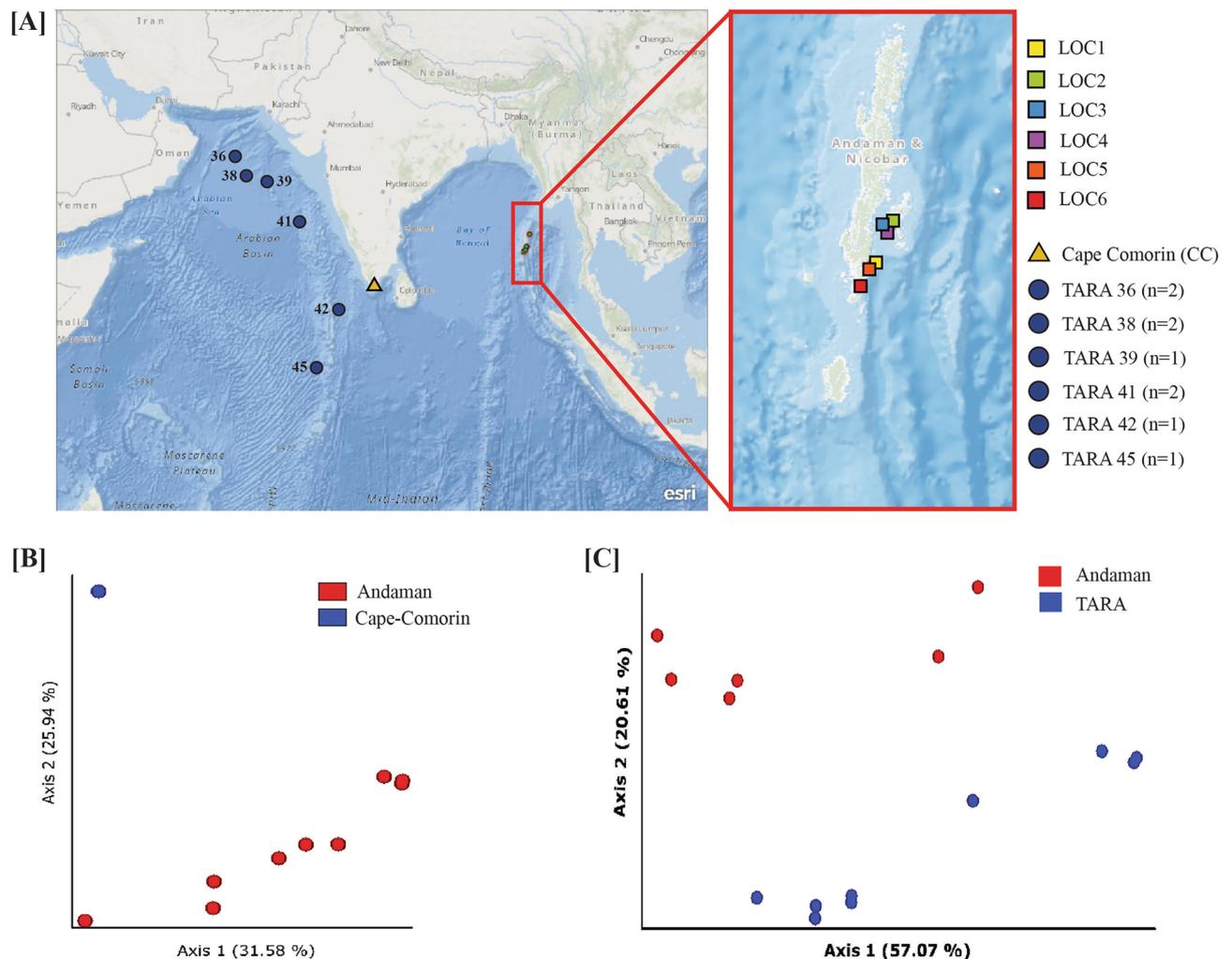


Figure 1. (A) The map shows locations (in colored rectangles) of coastal water sample collection sites from Andaman and Nicobar Islands. Blue colored circles represent Tara Oceans sampling stations in the Indian Ocean region for comparative analysis. (B) Principal coordinates analysis based on Bray–Curtis distance between 16S rRNA amplicon samples. Sample from Cape-Comorin clusters away from Andaman samples. (C) Principal coordinates analysis based on Bray–Curtis distance using microbial species abundance in each sample revealed a separation between Tara Oceans and Andaman samples; PCo-1 explained 57.07% of the variation.

Abundance of *Synechococcus* species in coastal waters of Andaman Islands

Taxonomic annotation of ASVs resulted in the identification of 59 phyla, 321 families, and 574 genera. Proteobacteria and Cyanobacteria (37% and 35% average relative abundance) showed the highest abundance followed by Firmicutes, Actinobacteria, Bacteroidetes, Acidobacteria, and Gemmatimonadetes in all samples including Andaman and CC. All these phyla showed > 1% abundance in all samples. At the family level, *Synechococcaceae*, *Bacillaceae*, and *Rhodobacteriaceae* (relative abundance 38.19, 11.97, and 5.37%) were the most abundant families followed by *Pelagibacteriaceae*, *Oceanospirallaceae*, *Alteromonadaceae*, *Methylobacteriaceae*, *Propionibacteriaceae*, and *Nocardioideaceae*. *Synechococcus* showed the highest abundance in Andaman samples with an average relative abundance of 37%. A core microbiome analysis was performed considering genera present in all samples with > 1% average relative abundance. *Synechococcus*, *Bacillus*, *Marinomonas*, *Alteromonas*, *Methylobacterium*, and *Pseudoalteromonas* were identified as core genera in Andaman samples. *Lutimonas* was found to be the most abundant genus in the CC sample, followed by *Anaerospira*, *Tenacibaculum*, *Thalassobius*, *Acidaminobacter*, *Nisaea*, *Propionibacterium*, and *Corynebacterium* (Fig. 2A, Supplementary Fig. 3A, B, C).

Further, we validated the genus-level microbial composition by taxonomic annotation of metagenomic reads using Kaiju (Supplementary Fig. 4A). *Synechococcus* (13.3%), *Haloparvum*, *Methylobacterium*, *Halorubrum*, *Candidatus*, *Marinomonas*, and *Alteromonas* were identified as core bacterial genera in Andaman samples. Whereas, *Prochlorococcus* (7.25%), *Candidatus Pelagibacter*, *Alteromonas*, and *Synechococcus* were detected as core bacterial genera in open ocean samples of Tara Oceans Consortium (Fig. 2B, C, Supplementary Table 5). The bacterial species abundance in coastal and open ocean samples analyzed using Kraken2 revealed *Methylobacterium populi*, *Synechococcus* sp. LTW-R, *Synechococcus* sp. TAK9802, *Synechococcus* sp. M16.1 and *Alteromonas macleodii* as the most abundant species with their average relative abundance ranging between 3.9 and 4.8% in coastal samples

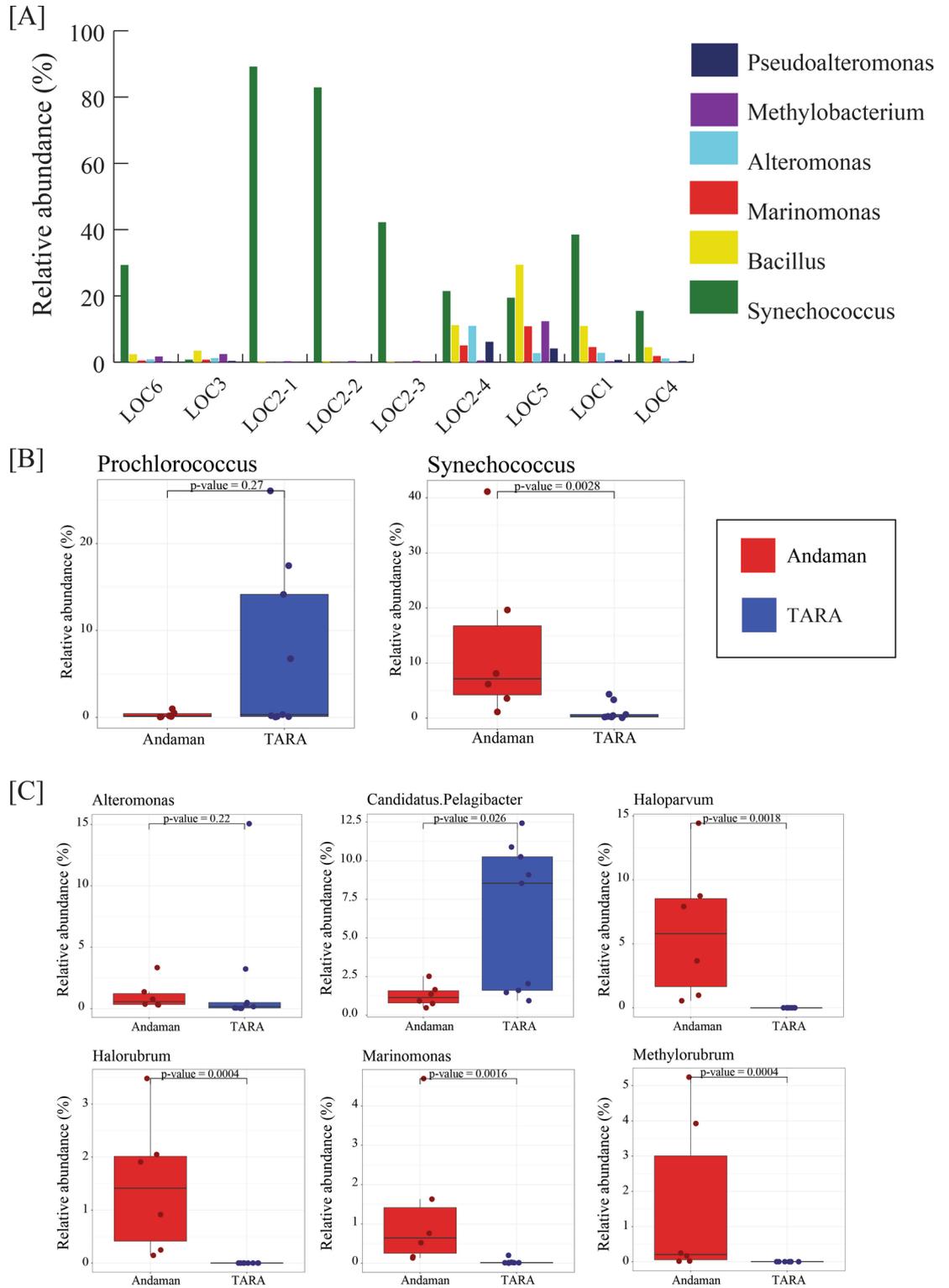


Figure 2. (A) Relative abundance of most abundant genera in 16S rRNA amplicon samples from coastal waters. (B) Figure shows the relative abundance of *Synechococcus* and *Prochlorococcus* in metagenomic data of open oceans (Tara Oceans) and coastal water (Andaman) samples. (C) Figure shows relative abundance of core genera (> 1% average abundance) identified in coastal (Andaman) and open ocean (Tara Oceans) samples. The relative abundance of all core genera (except *Alteromonas*) was significantly different in coastal and open oceans.

from Andaman. Whereas, *Prochlorococcus marinus*, showed the highest abundance (average relative abundance 12.5%) in open ocean samples (Supplementary Table 5).

Analysis of the top 20 microbial species in coastal and open ocean samples indicated a higher number of species (10 out of 20) from *Synechococcus* in coastal samples (Fig. 3A, Supplementary Fig. 4B, Supplementary Table 5). On the other hand, the top 20 microbial species in open ocean samples were assigned to *Prochlorococcus* (five sp.), *Candidatus Pelagibacter* (six sp.), *Alteromonas* (four sp.), *Synechococcus* (four sp.), and *Pseudomonas resinovorans* (Fig. 3B, Supplementary Fig. 4C, Supplementary Table 5). Briefly, a higher abundance and richness of *Synechococcus* species in coastal water samples and a higher abundance of *Prochlorococcus* species were observed in open ocean samples.

Differentially abundant microbial genera and species in coastal water and open ocean surface water samples

We detected 51 microbial genera exhibiting differential abundance between coastal water and open ocean surface water using LEfSe and Boruta. Out of 51 genera, 30 were Archaeobacterial halophiles from class *Halobacteria*.

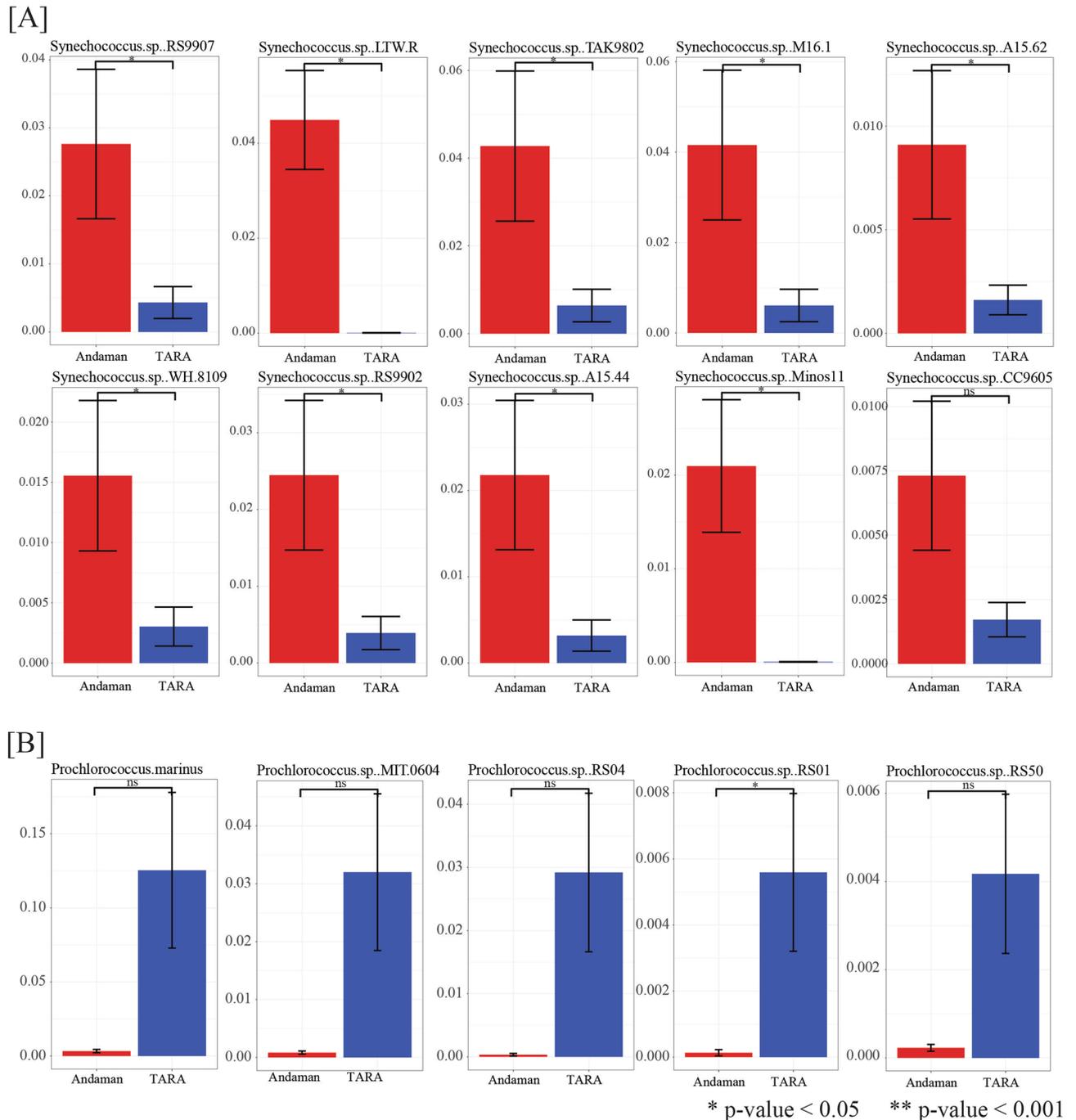


Figure 3. (A) 10 *Synechococcus* and (B) five *Prochlorococcus* species detected in most abundant 20 species in metagenomic data of coastal (Andaman) and open ocean (Tara Oceans) samples.

These genera were highly abundant in coastal water samples. Many of these genera have been isolated from hypersaline environments, marine sediments, and soil (Supplementary Table 6). The *Hyphomonas* (class Alphaproteobacteria) was the only genus among the differentially abundant genera that exhibited higher abundance in the open ocean samples. This bacterium is a dimorphic, prosthecate bacteria ubiquitous in the marine environment⁵⁹ (Supplementary Fig. 5, Supplementary Table 6).

Analysis using Boruta and LEfSe identified 47 differentially abundant microbial species (from kraken2 analysis) in coastal water and open ocean surface water. Out of these, 41 species showed significant abundance in coastal waters, and notably, 32 of these (~69.56%) were identified as salt-tolerant species (Supplementary Fig. 6, Supplementary Table 6). Five viruses of *Synechococcus* and *Prochlorococcus* (Supplementary Table 6) were found to be differentially abundant in open oceans based on the abundance values obtained using kraken2. Marine viruses have active roles in shaping the physiology and diversity of picocyanobacterial species (Supplementary Fig. 6)⁶⁰.

Higher diversity of *Synechococcus* genomes in coastal waters compared to *Prochlorococcus* in open ocean water samples

To further explore the genome-level composition of the most abundant genera in coastal and open ocean samples, we constructed genome databases for both *Prochlorococcus* and *Synechococcus* by retrieving 1100 and 405 genomes, respectively from NCBI GenBank database (Fig. 4A, B). The average genome size of *Prochlorococcus* is smaller than *Synechococcus* (p value = $1.750e-88$) (Fig. 4B). Whereas, the variation in genome size of *Synechococcus* is significantly higher than *Prochlorococcus*. We calculated the abundance of all 1505 genomes in all 15 samples and used the top 100 genomes in both datasets for further analysis. In the resulting data, 81 out of the top 100 genomes in the open ocean samples were from the *Prochlorococcus* genus. Similarly, 85 out of the top 100 genomes in coastal water samples were from the *Synechococcus* genus (Supplementary Table 7). Among the top 20 genomes in the data, 11 *Synechococcus* genomes were abundant in coastal samples (Fig. 4C, Supplementary Table 7) and others were abundant in the open ocean group.

We also constructed a phylogenetic tree using 16S rRNA sequences predicted from the most abundant *Synechococcus* ($n=45$ out of 85) and *Prochlorococcus* ($n=70$ out of 81) genomes in coastal and open ocean microbiomes. Higher phylogenetic diversity was observed in a phylogenetic tree constructed using *Synechococcus* genomes, whereas the most abundant *Prochlorococcus* genomes from open ocean samples formed only two major clades (Supplementary Fig. 7).

Non-redundant gene set of most abundant *Synechococcus* and *Prochlorococcus* genomes and their functional insights

To evaluate the functional potential of top abundant *Synechococcus* ($n=85$) and *Prochlorococcus* ($n=81$) genomes in coastal and open ocean waters microbiome, we constructed non-redundant gene sets for each genus by predicting genes from the top abundant genomes (see methods). *Synechococcus* and *Prochlorococcus* gene sets contained 137,371 and 88,718 non-redundant genes respectively. The difference in the number of genes in both datasets reflects the differences in the respective genome sizes of the two cyanobacteria.

We also analyzed the highly abundant functions (in terms of KOs) in *Prochlorococcus* and *Synechococcus* gene sets. Highly enriched KOs in both genera were involved in photoinduced damage repair and oxidative stress (*ftsH*, *phrB*, *gst*), ABC type transporter proteins, chlorophyll biosynthesis-related protein (K00218), carbohydrate metabolism (*galE*, *rfdD*), amino acid metabolism (K01802) and transcription (Supplementary Table 8). Differentially represented functions (KOs) in both gene sets were evaluated by the ratio of the number of genes assigned to each KO in *Synechococcus* to the number of genes assigned to each KO in *Prochlorococcus* and vice versa (Syn/Pro and Pro/Syn ratio) (Supplementary Table 8).

Identification and annotation of orthologous gene clusters in non-redundant gene sets of the most abundant *Synechococcus* and *Prochlorococcus* genomes

OrthoFinder and OrthoVenn2 were used to examine the orthogroups present in *Synechococcus* and *Prochlorococcus* gene sets. Analysis using OrthoVenn2 at 100% identity cut-off resulted in 8075 orthologous gene clusters from *Synechococcus* and 4,818 orthologous gene clusters from *Prochlorococcus*. Of these, 5304 and 2047 unique orthologous gene clusters were identified in *Synechococcus* and *Prochlorococcus*, respectively (Fig. 4D). Further validation using gene sets constructed at a 90% identity threshold identified 5890 and 3605 orthologous genes (total of 7347) clusters in *Synechococcus* and *Prochlorococcus*, respectively (Fig. 4E). At the 90% identity threshold, 3742 and 1457 unique orthologous gene clusters were identified in *Synechococcus* and *Prochlorococcus*, respectively. The longest genes from each cluster were selected as representative and annotated based on a homology search against the nr database using DIAMOND. Out of 7347 genes, 4307 (58.62%) representative genes were annotated using the criteria of $\geq 90\%$ identity and $\geq 80\%$ query coverage (Supplementary Tables 9, 10).

Confirmatory analysis using OrthoFinder revealed 9295 orthogroups (containing at least two genes per cluster) within *Synechococcus* (7263 orthogroups) and *Prochlorococcus* (4007 orthogroups). Of these, only a fraction, 64.20% (5968 out of 9295) were annotated using the nr database. In particular, we identified 5288 unique orthogroups in *Synechococcus* and 2032 in *Prochlorococcus*. A higher number of unique gene clusters present in *Synechococcus* genera indicates the potential of *Synechococcus* genomes to carry out diverse and distinct functions compared to *Prochlorococcus* genomes (Supplementary Tables 9, 10). A large part of these clusters was annotated as hypothetical proteins using DIAMOND-BLAST against the nr database.

We also examined the presence of photosynthesis-related proteins in both *Synechococcus* and *Prochlorococcus* orthologous gene sets using the KEGG database. Photosynthesis-related proteins are represented as three major groups in the KEGG database namely, photosystem and electron transport system, antenna proteins, and

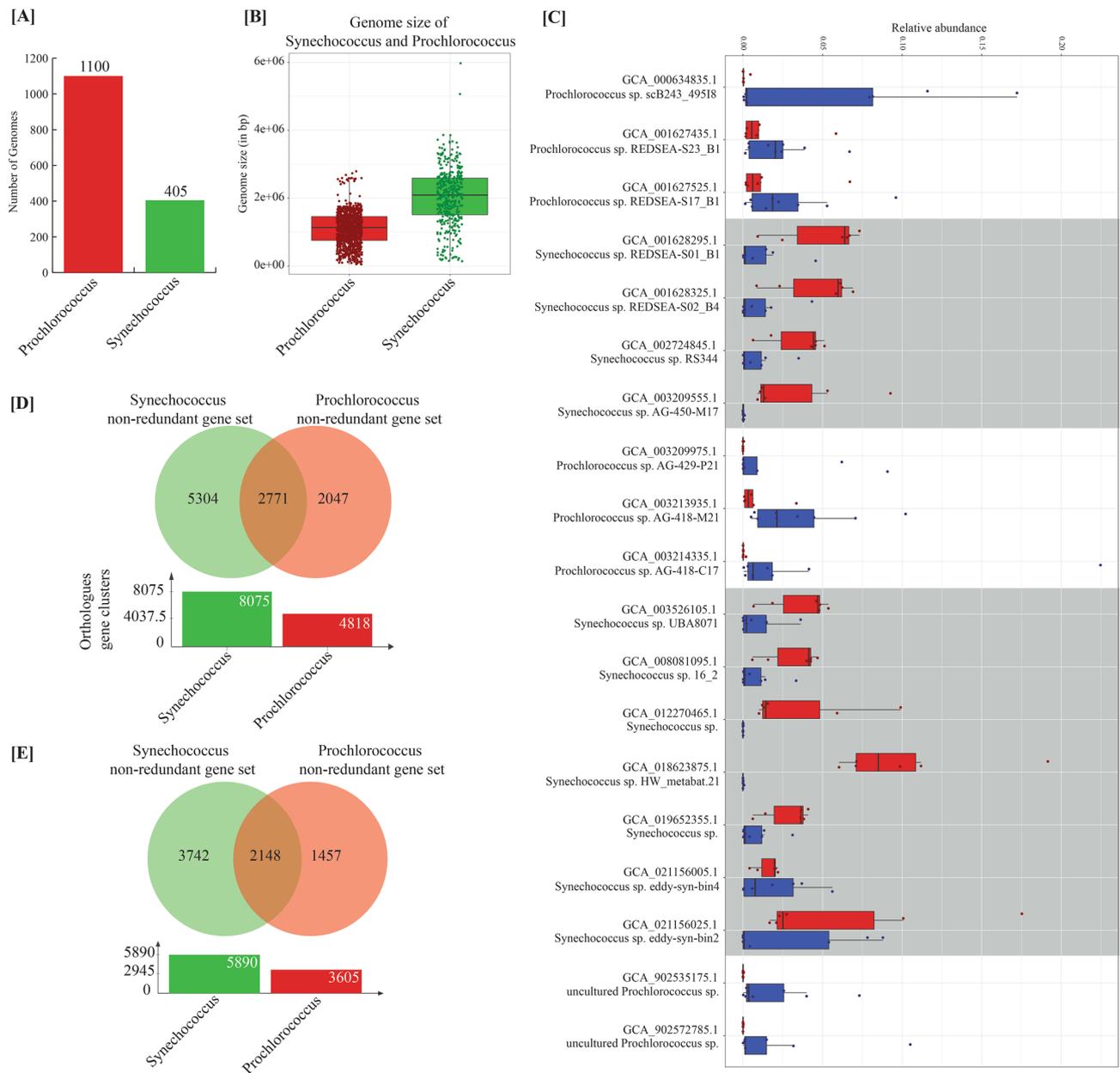


Figure 4. (A) Number of *Synechococcus* and *Prochlorococcus* genomes retrieved from NCBI Genbank database for analysis. (B) Box-plot representing genome size of *Synechococcus* and *Prochlorococcus* genomes used in the analysis. (C) Top abundant *Synechococcus* and *Prochlorococcus* genomes in Andaman and Tara Oceans samples based on alignment of metagenomic reads to genomes. Genomes highlighted in Grey were abundant in Andaman coastal samples. (D) Venn-diagram of orthologous gene clusters of *Synechococcus* and *Prochlorococcus*. The gene sets of *Synechococcus* and *Prochlorococcus* used for this analysis were constructed by clustering of ORFs predicted from both genera at 100% amino acid identity. (E) Venn-diagram of orthologous gene clusters of *Synechococcus* and *Prochlorococcus*. The gene sets of *Synechococcus* and *Prochlorococcus* used for this analysis were constructed by clustering of ORFs predicted from both genera at 90% amino acid identity.

anoxygenic photosystem. Among the orthologous gene sets identified via OrthoFinder, we found 28 KOs associated with the photosystem and electron transport system within *Synechococcus*. In contrast, *Prochlorococcus* exhibited only 11 KOs related to these functions. Similarly, 18 KOs in orthologous gene sets (OrthoFinder) of *Synechococcus* and 2 KOs in orthologous gene sets (OrthoFinder) of *Prochlorococcus*, were mapped to antenna proteins.

Updated Ocean microbiome gene catalog

Sample wise assembly and gene prediction from Andaman samples resulted in 113,410 non redundant genes of > 100 bp nucleotides length (Supplementary Table 11). We updated the ocean microbiome gene catalog by adding 113,410 genes to the Ocean Microbial Reference Gene Catalog (OMRGC). The updated non-redundant

catalog contained 39,732,487 genes. A reduction in the number of genes in the updated gene catalog indicates the redundancy in OMRGC. The updated gene catalog contains 93,863 genes from the Andaman region. Out of these, 93,172 (99.26%) genes were part of unique clusters and are not covered in OMRGC. Further analysis of 691 genes that were clustered with genes from OMRGC resulted in the identification of 298 representative genes from Andaman that replaced OMRGC genes with more than two-fold their length. Only a limited number of newly added genes could be annotated using eggNOG-mapper v2. Only 53,934 (57.46%) genes out of 93,863 genes were annotated using DIAMOND against the nr database with $\geq 90\%$ identity and $\geq 80\%$ query coverage (Supplementary Table s 12). Further, only 2688 genes (2.86%) were assigned molecular functions in KEGG, 45 (0.048%) were assigned by CAZy, 1608 (1.71%) were assigned to EC classes, and 3919 (4.17%) to COG classes. Thus, a total of 53,934 (57.46%) genes out of 93,863 genes were successfully annotated with a specific function. The presence of a large number of unannotated proteins might represent new and previously unknown functional diversity from coastal surface waters. We looked for the KOs involved in photosynthesis and found 26 genes annotated as photosynthetic proteins in genes newly contributed from our study. Similarly, 12 genes were involved in the betaine biosynthesis pathway, which is a compatible solute and provides the cells with adaptations of osmotic stress.

Functional role of ocean microbiome in the coastal and open ocean region

We calculated the abundance of genes from the updated ocean microbiome gene catalog in all coastal and open ocean samples. Out of a total of 39,732,487 genes, 5,493,722 (13.82%) were identified in at least one surface water sample. A total of 2,250,966 (40.97%) genes were assigned with KEGG Orthologs (KO) using the KEGG database. Functional analysis was conducted on 1,867,938 genes detected in over 20% of the samples (i.e., at least 3 out of 15 samples) and exhibiting a relative abundance exceeding 0.05% per sample. Alpha diversity based on Shannon evenness and observed KOs in coastal and open ocean samples were also calculated (Supplementary Table 13). Beta diversity analysis using Jaccard distance indicated distinct clustering of coastal and open ocean samples (Fig. 5A).

Functions related to the transport of various substrates (7 transporters out of top 20 abundant functions), which include transporters for the acquisition of iron and copper (*ABC.CD.P*, *TC.FEV.OM*), amino acid transporters (*livK*), genes for transport of ammonium (*amt*) and transporters for the acquisition of sugars and branched amino acids were found in coastal water samples. Genes involved in DNA replication and repair (*rpoC*, *phrB*), amino acid metabolism (*glnA*, *gcvT*, *DMGDH*), photosynthesis (*psbA*), and photoreactivation after UV damage (*phrB*) also showed higher abundance in coastal samples. Genes involved in osmoregulation and oxidative stress (*DMGDH*, *GST*, sugar transporters, *ftsH*) were also abundant in coastal waters.

Open ocean samples showed a higher abundance of genes involved in DNA replication and repair (*polA*, *uvrD*, *MGME1*), purine nucleotide biosynthesis (*nrdA*, *nodU*), photosynthesis and electron transport (*petF*), heat shock proteins, and chaperone proteins (*groES*, *PEO1*, *groEL*, *ibpA*) and complex carbohydrate biosynthesis related functions (*UGDH*, *TSTA3*, *tagD*, *galE*). Iron acquisition transporter (*TC.FEV.OM*) was also among the most abundant functions in the open ocean microbiome.

Genes involved in photosynthesis and oxidative phosphorylation (*apcE*, *atpB*, *coxC*, *ctaE*, *cpcC*, *cpeA*) were differentially abundant in Andaman coastal water samples. Many transporters related to peptide transport (*ABC.PE.S* & *ABC.PE.P1*, *aapQ*), and polyamine uptake were also differentially abundant which serve as carbon, nitrogen, and/or energy source for marine bacterioplankton. Many other differentially abundant genes in coastal samples were transporters for sugar transport (fructose transport system, multiple sugar transport system ATP-binding protein) and Na^+/H^+ antiporter involved in pH and ionic homeostasis. Other differentially abundant genes related to amino acid biosynthesis pathways, stress response, purine nucleotide synthesis (*pncA*, *pncB*, *gpt*), metal stress (*chrA*, K07240), NAD biosynthesis, and carbohydrate metabolism genes, hydrocarbon degradation (*cntA*) and dehalogenases (*queG*) were also found at higher abundance in coastal samples than open ocean samples. Differentially abundant functions in open ocean samples include DNA replication and repair, carbohydrate biosynthesis (*galE*, *glf*, *hddA*), and iron homeostasis (*iscA*, *erpA*).

Carbohydrate metabolizing genes analysis

A higher abundance of carbohydrate metabolizing genes in coastal and open ocean samples identified by KEGG analysis motivated an in-depth investigation of carbohydrate metabolizing enzyme families using the carbohydrate-active enzyme (CAZy) database. The updated gene catalog revealed a total of 58,207 CAZy genes. Differentially abundant CAZy genes in coastal and open ocean water were detected using Boruta and LEfSe. This analysis showed higher abundance of complex carbohydrate hydrolysis related functions like beta-NAGsidas, and glycoside hydrolases with varying substrate specificities in coastal waters. The abundance of genes involved in carbohydrate degradation and hydrolysis functions indicates more heterotrophic activity in coastal waters. In addition, a higher abundance of trehalose phosphate synthase (GT20) was also observed in coastal samples compared to open ocean water. Whereas open ocean samples showed a differential abundance of complex CAZy families transferases catalyzing various complex polysaccharide synthesis (GT25, GT31, GT15, GT47, etc.). Pectin degradation-related CAZy families (PL4, PL11) were also found to be abundant in open ocean samples (Fig. 5C).

Antibiotic resistance genes analysis

To examine the impact of the higher human activity in coastal regions, we analyzed the composition of antibiotic resistance genes in coastal waters and compared it with open ocean samples. Antibiotic resistance genes were identified from 5,493,722 genes represented in the updated gene catalog using RGI. Differentially abundant Antibiotic Resistance Ontology (ARO), AMR gene families, and drug classes were identified using LEfSe and Boruta (Fig. 6). Genes categorized (by RGI) as conferring resistance to aminoglycoside, beta-lactam (monobactam;

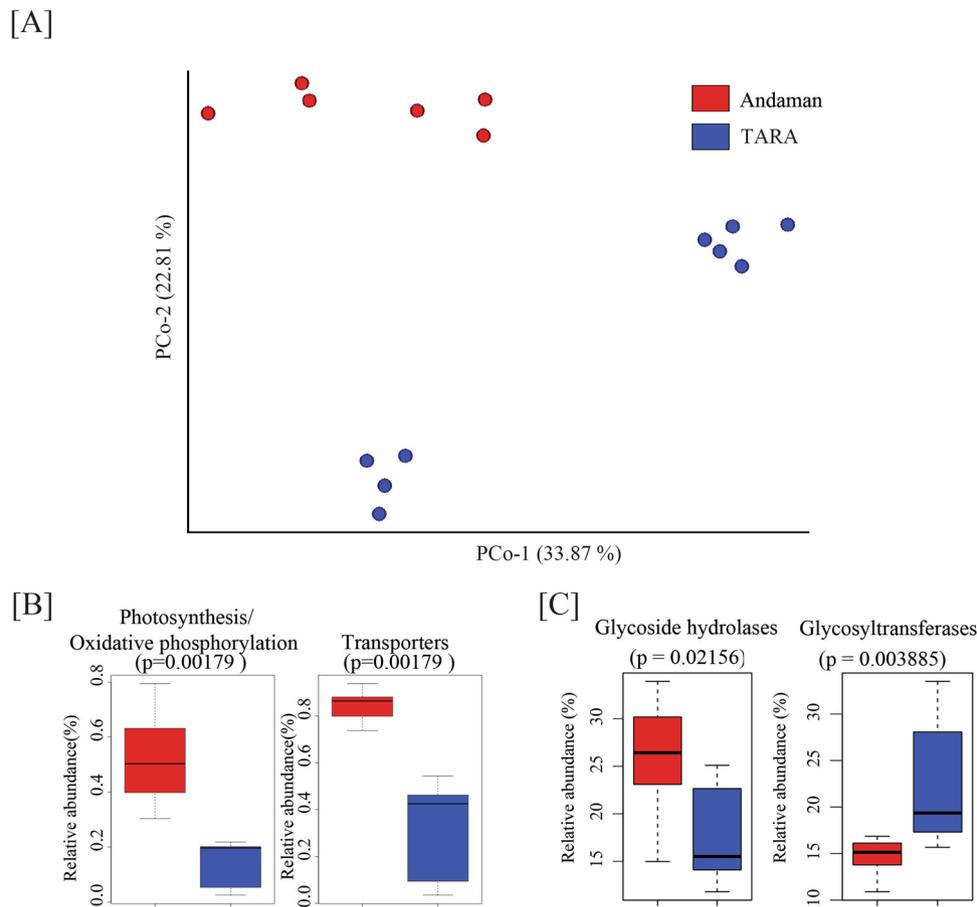


Figure 5. (A) Principal coordinates analysis based on relative abundance of KOs of coastal and open ocean samples. (B) Relative abundance of photosynthesis/oxidative phosphorylation related functions and transporters in Andaman and Tara Oceans samples. (C) Relative abundance of Glycoside Hydrolase and Glycosyltransferase CAZy gene families in Andaman and Tara Oceans samples.

cephalosporin; penam; penem) and tetracycline antibiotics were identified as differentially abundant in coastal samples. Most of the differentially abundant AMR functions were in higher abundance in coastal waters compared to the open ocean. Analysis of AROs indicated a higher abundance of TEM-116, BES-1, APH(7'')-Ia, APH(3')-Ia, FomB which correspond to beta-lactam, and aminoglycoside antibiotic resistance in coastal waters. The top five differentially abundant AMR gene families in Andaman coastal waters include TEM beta-lactamase, BES beta-lactamase, APH(7''), ADC beta-lactamases (with undetermined carbapenemase activity), and major facilitator superfamily (MFS) antibiotic efflux pump (Fig. 6, Supplementary Table 14).

Discussion

The study of the microbial composition and interactions in specific ocean regions promises to enhance our understanding of the functional potential of ocean microbiomes and their roles in maintaining healthy marine ecosystems. While previous studies have focused on the open ocean, the microbial composition of coastal waters remains understudied. This study reveals the taxonomic and functional diversity of the coastal region of Andaman using the 16S rRNA gene amplicon and whole metagenomic data analysis. The comparative analysis of coastal water microbiome with samples from the open ocean region of marine pelagic biomes of ARAB and MONS retrieved from the Tara Oceans Consortium revealed key differences between the two ecosystems.

One of the main outcomes of this study is the construction of an updated ocean microbiome gene catalog by the inclusion of 93,172 unique genes from coastal samples that will help to unravel the yet unexplored functional diversity of the coastal waters in Andaman with potential biotechnological applications^{24,61}. Notably, the *Synechococcus* species was significantly abundant in Andaman coastal water compared to open ocean samples from the Tara Oceans Consortium, consistent with previous reports of their widespread distribution in surface waters⁶². As a photosynthetic bacterium, *Synechococcus* exhibits efficient adaptations to variable salinity, light intensity, and temperature⁶³. Certain strains of *Synechococcus* demonstrate distinct photosynthetic activity during the summer season, contrasting with other seasons⁶⁴. On the contrary, *Prochlorococcus* genomes showed higher abundance (not significant) in open ocean samples compared to the coastal waters of Andaman. *Prochlorococcus* thrives in sunlit and well-oxygenated, nutrient-poor, tropical, and subtropical waters^{62–64}.

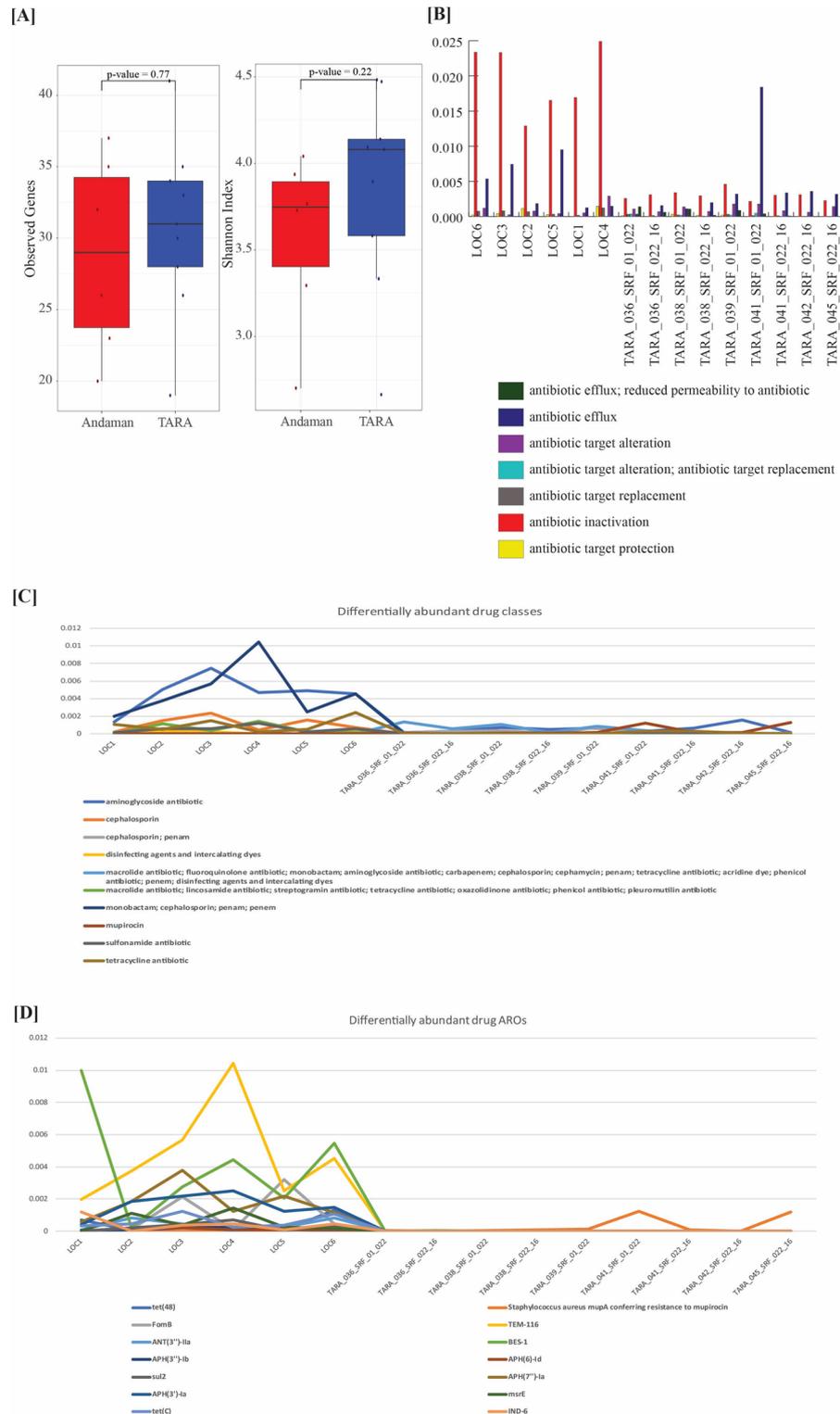


Figure 6. (A) Box plot based on alpha-diversity of Antibiotic Resistance Ontology (ARO) classified by RGI. The number of genes (observed AROs) and Shannon Index were calculated based on the proportion of differentially abundant AROs identified using LEfSe and Boruta. (B) Relative proportions of various Antibiotic resistance mechanisms detected in samples from coastal and open ocean samples. (C) Relative proportion of differentially abundant drug classes in which the resistance was detected in coastal and open ocean waters. (D) Relative proportions of Antibiotic Resistance Ontology (ARO) in each sample. All AROs were highly abundant in coastal water samples compared to the open ocean.

Synechococcus and *Prochlorococcus* differ in their light-harvesting apparatus, with *Synechococcus* utilizing chlorophyll a and *Prochlorococcus* relying on divinyl chlorophylls a and b⁶⁵. In-depth analysis constructing comprehensive genome databases of *Synechococcus* and *Prochlorococcus* revealed a higher abundance of genomes reconstructed from metagenomes (MAGs) in both coastal and open ocean waters indicating the yet under-explored diversity of *Synechococcus* and *Prochlorococcus* species. Also, *Synechococcus* genomes were significantly larger and more variable in size compared to *Prochlorococcus* genomes⁶⁵.

Analysis of non-redundant gene sets (nr gene sets) of highly abundant genomes of *Synechococcus* and *Prochlorococcus* revealed the functional roles of these species in coastal and open ocean waters. We observed a higher abundance of *psbA* gene, which encodes D1 protein of PSII, in coastal water samples. This gene is known to be highly expressed in *Synechococcus* sp. WH7803 under changing light conditions, indicating the adaptability of *Synechococcus* species to light stress in coastal waters⁶⁶. The presence of terrestrial material, algae, and suspended biomass in coastal waters leads to different light quality compared to open oceans and may influence the distribution of *Synechococcus* and *Prochlorococcus* species. The incidence of fewer pigment-related genes in the *Prochlorococcus* gene set signifies its adaptation to relatively stable light conditions prevalent in open ocean environments. Conversely, *Synechococcus* genomes showcase a broader spectrum of pigment types, likely an adaptive strategy to cope with varying light regimes. The diverse array of pigments in *Synechococcus* genomes from coastal waters equips them to acclimate to fluctuating light conditions and nutrient-rich environments. In contrast, *Prochlorococcus* exhibits limited functional diversity and thrives primarily in relatively uniform oligotrophic waters.

Coastal waters exhibit variable salinity ranges and the presence of coastal lagoons and salt marshes can result in hypersaline conditions. Differential abundance of halophile genera (*Haloarcula* and *Haloquadratum*) and compatible solute biosynthesis-related KOs (e.g. KO10112) in coastal waters indicated higher adaptations like osmotic adjustments in coastal microbial community^{67,68}. Several salt-tolerant, halophilic, or haloalkaliphilic bacterial/archaeal species were found to be differentially abundant in the coastal waters of Andaman. Various strains of differentially abundant species have been isolated from environments of high-salinity. (Supplementary Table 6). Specifically, the lack of water channel protein *AqpZ*, a channel that permits osmotically driven movement of water in both directions, and is involved in the osmoregulation and maintenance of cell turgor during volume expansion in rapidly growing cells, in almost all picoplanktonic cyanobacteria (with the exception of the coastal euryhaline strain *Synechococcus* sp. strain WH5701)⁶⁵ further highlights the possibility of adaptations like osmotic adjustments in coastal microbial communities.

Due to their susceptibility to anthropogenic activities and associated stress conditions, coastal waters create an environment conducive to the presence and accumulation of ARGs^{69,70}. In this study, the analysis of ARGs in coastal and open ocean water samples revealed the presence of 64 AMR gene families in the Andaman region, in which the genes conferring resistance to aminoglycoside, beta-lactam, and tetracycline antibiotics were highly abundant in coastal waters. We also observed NDM beta-lactamase in Andaman samples and five Indian Ocean samples (not significantly different). The NDM class (New Delhi metallo-beta-lactamase) of beta-lactamase enzymes hydrolyzes a wide range of beta-lactam antibiotics, including carbapenems. These antibiotics are critical for combating drug-resistant bacterial infections. Originally reported in India, the prevalence of NDM has expanded globally, posing a significant public health concern^{23,71,72}. Most of the differentially abundant ARGs showed higher abundance in coastal waters compared to the open ocean. A recent study analyzing globally distributed Tara Oceans data identified tetracycline and beta-lactam resistance as prevalent ARGs⁷³. Fosmidomycin, quinolone, and bacitracin ARGs were also reported to be higher in coastal biomes in Tara Oceans data. A recent study reported higher abundance of quinolone ARGs in coastal waters from China⁶⁹. Additionally, in the Indian Ocean Commission (IOC) study encompassing various island countries within the Indian Ocean, carbapenemase-producing *Enterobacteriaceae* and extended-spectrum beta-lactamase were identified as principal concerns for human and animal health⁷⁴.

In summary, this study revealed the microbiome composition of the coastal waters of Andaman Islands. Comparative analysis with selected open ocean samples from Tara Oceans consortium revealed differences between coastal and open ocean habitats. Increased abundance, prevalence of diverse species, and distinct functional composition of *Synechococcus* genus in coastal samples indicate their adaptability to variable environmental conditions and nutrient enrichment. A detailed analysis of abundant *Synechococcus* genomes revealed enrichment of photosynthesis pigment-related genes, suggesting their adaptation to the variable light conditions in coastal waters. The differential abundance of halophile species observed in coastal waters corroborates well with highly variable salinity in coastal areas. The inclusion of 93,172 unique genes from Andaman coastal water samples with a large number of unannotated genes in the updated gene catalog indicated the yet unexplored functional potential of these areas. This study also highlights the significance of monitoring coastal waters for antibiotic resistance due to human impacts and the potential risk of Horizontal Gene Transfer (HGT) of Antimicrobial resistance genes.

Even though this study was performed with a limited number of samples from a coastal region of a single geographical area, it represents the first metagenomic exploration of Indian coastal locations from the Andaman and Nicobar Islands. Considering the long coastline (~7500 km) of India with diverse environmental features, such as mangroves, coral reefs, seagrass beds, estuaries and lagoons, backwaters, etc., the knowledge and understanding of taxonomic and functional diversity in Indian coastal waters is much needed along with the studies to examine the impacts of anthropogenic activities on a much larger scale. Further studies based on longitudinal sampling with comprehensive environmental metadata from diverse habitats with different levels of anthropogenic impacts will provide more insights into microbial community variation due to high human activity. Exploring the microbial composition of coastal waters from diverse geographical regions will shed light on the underexplored diversity of *Synechococcus* genotypes.

Data availability

The raw sequencing data is available on NCBI BioProject database under study accession PRJNA822508 and PRJNA295549.

Received: 23 August 2023; Accepted: 28 January 2024

Published online: 01 February 2024

References

- Falkowski, P. Ocean Science: The power of plankton. *Nature* **483**, S17–S20 (2012).
- Zehr, J. P. Nitrogen fixation by marine cyanobacteria. *Trends Microbiol.* **19**, 162–173 (2011).
- Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**, 1034–1039 (2008).
- Ustick, L. J. *et al.* Metagenomic analysis reveals global-scale patterns of ocean nutrient limitation. *Science* **372**, 287–291 (2021).
- Hoffman, L. Marine cyanobacteria in tropical regions: Diversity and ecology. *Eur. J. Phycol.* **34**, 371–379 (1999).
- Partensky, F., Blanchot, J. & Vault, D. Differential distribution and ecology of *Prochlorococcus* and *Synechococcus* in oceanic waters: a review (1999).
- Hoffman, L. & Hoffmann, L. Marine cyanobacteria in tropical regions: Diversity and ecology. *Eur. J. Phycol.* <https://doi.org/10.1080/09670269910001736432> (2010).
- Biller, S. J., Berube, P. M., Lindell, D. & Chisholm, S. W. *Prochlorococcus*: The structure and function of collective diversity. *Nat. Rev. Microbiol.* **13**(1), 13–27. <https://doi.org/10.1038/nrmicro3378> (2015).
- Palenik, B. *et al.* Genome sequence of *Synechococcus* CC9311: Insights into adaptation to a coastal environment. *Proc. Natl. Acad. Sci.* **103**(36), 13555–13559 (2006).
- Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
- Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B. & Williamson, S. The Sorcerer II Global Ocean sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* **5**, 398 (2007).
- Díez, B. *et al.* Metagenomic analysis of the Indian Ocean picocyanobacterial community: Structure, potential function and evolution. *PLoS One* **11**, e0155757 (2016).
- Salazar, G. *et al.* Global diversity and biogeography of deep-sea pelagic prokaryotes. *ISME J.* **10**, 596–608 (2015).
- Wang, Y. *et al.* Metagenomic analysis reveals microbial community structure and metabolic potential for nitrogen acquisition in the oligotrophic surface water of the Indian Ocean. *Front. Microbiol.* **12**, 229 (2021).
- Fernandes, G. L., Shenoy, B. D. & Damare, S. R. Diversity of bacterial community in the oxygen minimum zones of Arabian Sea and Bay of Bengal as deduced by illumina sequencing. *Front. Microbiol.* **10**, 3153 (2020).
- Angelova, A. G., Ellis, G. A., Wijesekera, H. W. & Vora, G. J. Microbial composition and variability of natural marine planktonic and biofouling communities from the Bay of Bengal. *Front. Microbiol.* **10**, 2738 (2019).
- Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, 1–10 (2015).
- Jaiswal, S. K., Saxena, R., Mittal, P., Gupta, A. & Sharma, V. K. Draft genome sequence of *Pseudomonas hussainii* strain MB3, a denitrifying aerobic bacterium isolated from the rhizospheric region of mangrove trees in the Andaman Islands, India. *Genome Announc.* **5**, 10 (2017).
- Trevathan-Tackett, S. M. *et al.* A horizon scan of priorities for coastal marine microbiome research. *Nat. Ecol. Evol.* **3**, 1509–1520 (2019).
- Su, H. *et al.* High diversity of β -glucosidase-producing bacteria and their genes associated with scleractinian corals. *Int. J. Mol. Sci.* **22**, 3523 (2021).
- Zhang, X. X., Zhang, T. & Fang, H. H. P. Antibiotic resistance genes in water environment. *Appl. Microbiol. Biotechnol.* **82**, 397–414 (2009).
- Zheng, D. *et al.* A systematic review of antibiotics and antibiotic resistance genes in estuarine and coastal environments. *Sci. Total Environ.* **777**, 146009 (2021).
- Mittal, P., Prasoodanan, P. V., Dhakan, D. B., Kumar, S. & Sharma, V. K. Metagenome of a polluted river reveals a reservoir of metabolic and antibiotic resistance genes. *Environ. Microbiomes* **14**, 1–12 (2019).
- Chaudhary, N., Gupta, A., Gupta, S. & Sharma, V. K. BioFuelDB: A database and prediction server of enzymes involved in biofuels production. *PeerJ* **2017**, e3497 (2017).
- Sharma, A. K., Kumar, S. K. H., Dhakan, D. B. & Sharma, V. K. Prediction of peptidoglycan hydrolases- a new class of antibacterial proteins. *BMC Genomics* **17**, 411 (2016).
- Saxena, R. *et al.* Metagenomic analysis of hot springs in central India reveals hydrocarbon degrading thermophiles and pathways essential for survival in extreme environments. *Front. Microbiol.* **7**, 2123 (2017).
- Biller, S. J. *et al.* Marine microbial metagenomes sampled across space and time. *Sci. Data* **5**, 1–7 (2018).
- Janda, J. M. & Abbott, S. L. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *J. Clin. Microbiol.* **45**, 2761–2764 (2007).
- Chaudhary, N., Sharma, A. K., Agarwal, P., Gupta, A. & Sharma, V. K. 16S classifier: A tool for fast and accurate taxonomic classification of 16S rRNA hypervariable regions in metagenomic datasets. *PLoS One* **10**, e0116106 (2015).
- Patel, R. K. & Jain, M. NGS QC toolkit: A toolkit for quality control of next generation sequencing data. *PLoS One* **7**, e30619 (2012).
- Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing ReadsHDGV. *Adv. Environ. Biol.* **7**, 10–12 (2011).
- Magoč, T. & Salzberg, S. L. FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
- Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
- Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
- Desantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–5072 (2006).
- Luo, R. *et al.* SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience* <https://doi.org/10.1186/2047-217X-1-18> (2012).
- Menzel, P., Ng, K. L. & Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* <https://doi.org/10.1038/ncomms11257> (2016).
- Wood, D. E., Lu, J. & Langmead, B. DerrImproved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 1–13 (2019).
- Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: Estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).
- Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* **38**, e132–e132 (2010).

41. Li, W. & Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
42. Li, R. *et al.* SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).
43. Wang, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).
44. Dhakan, D. B. *et al.* The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. *Gigascience* **8**, 1–20 (2019).
45. Sharma, A. K., Gupta, A., Kumar, S., Dhakan, D. B. & Sharma, V. K. Woods: A fast and accurate functional annotator and classifier of genomic and metagenomic sequences. *Genomics* **106**, 1–6 (2015).
46. Huerta-Cepas, J. *et al.* Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
47. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
48. Cantalapiedra, C. P., Hern Andez-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
49. Huerta-Cepas, J. *et al.* eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
50. Alcock, B. P. *et al.* CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **48**, D517–D525 (2020).
51. Hyatt, D. *et al.* Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 1–11 (2010).
52. Wang, Y., Coleman-Derr, D., Chen, G. & Gu, Y. Q. OrthoVenn: A web server for genome wide comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* **43**, W78–W84 (2015).
53. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14 (2019).
54. Seemann, T. & Booth, T. Barrnap: Basic rapid ribosomal RNA predictor. *GitHub Repos* (2018).
55. Tamura, K., Stecher, G. & Kumar, S. MEGA11: Molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* **38**, 3022–3027 (2021).
56. Kursa, M. B. & Rudnicki, W. R. Feature selection with the Boruta package. *J. Stat. Softw.* **36**, 1–13 (2010).
57. Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, 1–18 (2011).
58. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, 2016).
59. Abraham, W. Hyphomonas. In *Bergey's Manual of Systematics of Archaea and Bacteria* (ed. Dedysh, S. N.) 1–14 (Wiley, 2020).
60. Middelboe, M. & Brussaard, C. P. D. Marine viruses: Key players in marine ecosystems. *Viruses* **9**, 302 (2017).
61. Sharma, V. K., Kumar, N., Prakash, T. & Taylor, T. D. MetaBioME: A database to explore commercially useful enzymes in metagenomic datasets. *Nucleic Acids Res.* **38**, D468–D472 (2010).
62. Sohm, J. A. *et al.* Co-occurring *Synechococcus* ecotypes occupy four major oceanic regimes defined by temperature, macronutrients and iron. *ISME J.* **10**, 333–345 (2015).
63. Callieri, C. *Synechococcus* plasticity under environmental changes. *FEMS Microbiol. Lett.* **364**, 229 (2017).
64. Flombaum, P. *et al.* Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 9824–9829 (2013).
65. Scanlan, D. J. *et al.* Ecological genomics of marine picocyanobacteria. *Microbiol. Mol. Biol. Rev.* **73**, 249–299 (2009).
66. Garczarek, L. *et al.* Function and evolution of the psbA gene family in marine *Synechococcus*: *Synechococcus* sp. WH7803 as a case study. *ISME J.* **2**, 937–953 (2008).
67. Klähn, S. & Hagemann, M. Compatible solute biosynthesis in cyanobacteria. *Environ. Microbiol.* **13**, 551–562 (2011).
68. Liu, H. *et al.* Complete genome sequence of *Haloarcula hispanica*, a Model haloarchaeon for studying genetics, metabolism, and virus-host interaction. *J. Bacteriol.* **193**, 6086–6087 (2011).
69. Lu, J. *et al.* Occurrence and spatial distribution of antibiotic resistance genes in the Bohai Sea and Yellow Sea areas, China. *Environ. Pollut.* **252**, 450–460 (2019).
70. Zhu, Y. G. *et al.* Continental-scale pollution of estuaries with antibiotic resistance genes. *Nat. Microbiol.* **24**(2), 1–7 (2017).
71. Khan, A. U., Maryam, L. & Zarrilli, R. Structure, Genetics and Worldwide Spread of New Delhi Metallo- β -lactamase (NDM): A threat to public health. *BMC Microbiol.* **17**, 1–12 (2017).
72. Kumarasamy, K. K. *et al.* Emergence of a new antibiotic resistance mechanism in India, Pakistan, and the UK: A molecular, biological, and epidemiological study. *Lancet Infect. Dis.* **10**, 597–602 (2010).
73. 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* **9**, 1–12 (2020).
74. Gay, N. *et al.* Review of antibiotic resistance in the Indian Ocean commission: A human and animal health issue. *Front. Public Health* **5**, 162 (2017).

Acknowledgements

The authors express gratitude to the NGS facility at IISER Bhopal for facilitating the sequencing experiments. R.S. and V.P.P.K. thank DST-INSPIRE for their research fellowship funding. S.K. and D.B.D. thank the University Grants Commission (UGC), India for the research fellowship funding. The authors sincerely thank the Principal Chief Conservator of Forests (PCCF), Andaman & Nicobar Islands, for granting permission to collect samples.

Author contributions

V.K.S. conceived the work and participated in the design of the study. R.S. performed sample collection, processing, library preparation, and sequencing work with the help of D.B.D. and P.W. V.P.P.K. and S.K. designed the computational analysis framework with inputs from V.K.S. V.P.P.K. and S.K. carried out all metagenomic data and statistical analysis, interpretation of results, and prepared the first draft of the manuscript under the supervision of V.K.S. All authors have read and approved the final manuscript.

Funding

This work was supported by the Centre for Research on Environment and Sustainable Technologies (CREST) and intramural funding at IISER Bhopal, Madhya Pradesh, India.

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/s41598-024-53190-1>.

Correspondence and requests for materials should be addressed to V.K.S.

Reprints and permissions information is available at www.nature.com/reprints.

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024