# Meta-analysis cum machine learning approaches address the structure and biogeochemical potential of marine copepods associated bacteriobiome

Balamurugan Sadaiappan[1, +], Chinnamani PrasannaKumar[1, +], Uthara Nambiar V[1], Mahendran Subramanian[2, 3], Manguesh U Gauns[1, *]

[1]Plankton Ecology Lab, Biological Oceanography Division, CSIR-National Institute of Oceanography, Dona Paula, Panaji-403004, Goa, India.

[2]Department of Bioengineering and Department of Computing, Imperial College London, South Kensington- SW72AZ, London, United Kingdom.

[3]Faraday-Fleming Laboratory, London, W148TL, United Kingdom.

Corresponding author: Manguesh U Gauns (gmangesh@nio.org)

+Equal contribution

**Abstract**

Copepods are the dominant members of the zooplankton community and the most abundant form of life. It is imperative to obtain insights into the Copepod Associated Bacteriobiomes (CAB) to identify specific bacterial taxa associated within a copepod and to understand how they vary between different copepods. Analysing the potential genes within the CAB may reveal their intrinsic role in the biogeochemical cycles. For this, machine-learning models and PICRUSt2 analysis were deployed to analyse 16S rDNA gene sequences (~16.5 million reads) of CAB belonging to five different copepod genera viz., *Acartia* spp., *Calanus* spp., *Centropages* sp., *Temora* spp., and *Pleuromamma* spp. Overall, we predict 50 sub-OTUs (Gradient Boosting Classifier) as important s-OTUs in five copepod genera. Among these, 15 s-OTUs were predicted as important s-OTUs in *Calanus* spp. and 20 s-OTUs as important s-OTUs in *Pleuromamma* spp. Four bacterial genera *Acinetobacter johnsonii*, *Phaeobacter, Vibrio shilonii* and Piscirickettsiaceae were identified as important s-OTUs in *Calanus* spp., and bacterial genera *Marinobacter, Alteromonas, Desulfovibrio, Limnobacter, Sphingomonas, Methyloversatilis, Enhydrobacter* and Coribacteriaceae were predicted as important s-OTUs in *Pleuromamma* spp. for the first time. Our meta-analysis revealed that the CAB of *Pleuromamma* spp. had a high proportion of potential genes responsible for methanogenesis and nitrogen fixation, whereas CAB of *Temora* spp. had a high proportion of potential genes involved in assimilatory sulphate reduction, denitrification and cyanocobalamin synthesis. The CAB of *Pleuromamma* spp. and *Temora* spp. have potential genes accountable for iron transport.

**Keywords:** Important s-OTUs, ANCOM, PICRUSt2, Functional genes, Biogeochemical cycles, Cyanocobalamin synthesis.

## 1. Introduction

Copepods (Subphylum Crustacea; Class Hexanauplia; Subclass Copepoda) are an abundant and diverse group of zooplankton in the ocean [1, 2]. They play a key role in energy transfer within the pelagic food web [3]. They are also well-known for their wide-ranging and flexible feeding approaches [4]. Copepods, usually not more than a few millimetres in length, support a wide range of bacterial communities, both internally and externally (due to the release of organic and inorganic nutrients during feeding and excretion) [1-3]. Also, it is an already established fact that there is an exchange of bacterial communities between the copepods and the water-column due to their feeding behaviour [5, 6] and copepods transfer microbes from the photic zone up to the middle of the twilight zone [3, 7, 8]. The different environmental conditions between the surrounding water and copepods favour different bacterial communities [6, 7, 9].

However, feeding also changes the composition of bacterial communities in the copepod gut, i.e. high abundance of Rhodobacteraceae was reported in *Acartia* sp. with full gut than its starved counterparts [10]. Copepods have mutualistic associations with (Gammaproteobacteria) *Pseudoalteromonas* spp.. Also, Gammaproteobacteria was found to be more abundant in starved *Centropages* sp., *Acartia* sp. [10] and *Pleuromamma* sp. [11]. Meanwhile, a notable change was observed among bacterial communities between the diapause phase and actively feeding *Calanus finmarchicus* [2]. In a similar way,

Flavobacteriaceae was meagred in copepods during diapause and abundant in actively feeding its counterparts [2]. Datta et al. [2] reported that *Marinimicrobium* (Alteromonadaceae) was relatively more abundant in deep-dwelling copepods than its shallow counterparts and concluded that the copepods have inter-individual microbiome variations but the factors driving these variations are still unknown. From these early reports its well-known that bacterial communities associated with copepods vary on many factors, based on feeding, the difference in stages of life, body size and their vertical migration through the water column. Moreover, there may be a particular relationship or symbiotic and a natural core microbiome that depends necessarily not on the food, but on the host environment [10]. Herein, the terminology 'bacteriobiome' means the total bacterial composition inhabiting in a specific biological niche (for example, copepods), including their genomic content and metabolic products [12]. It is a well-known fact that host-associated microbial communities remain essential for maintaining any ecosystems, and any variation in these communities can be unfavourable. So, studying the specific bacterial taxa associated with copepods and its variations as well as analysing the potential genes within the CAB will help us in understanding their role in the host health, marine food web and biogeochemical cycles.

Until now, only a few studies have sought to find the core-bacteria associated with the copepods using their clustering patterns [2] and presence/absence data [1]. From these studies, about eight bacterial orders, such as *Actinomycetales, Bacillales, Flavobacteriales, Lactobacillales, Pseudomonadales Rhizobiales* and *Vibrionales*, were found as core members in *Pleuromamma* spp. [1], whereas the phylum Proteobacteria was identified as core OTUs along with Actinobacteria and Bacteroidetes in *Calanus finmarchicus* [2].

Moreover, the gut of copepods has acidic pH and different oxygen gradient from the anal opening to the metasome region. This may influence certain groups of bacteria to colonise within the copepods. These bacterial communities could be specialised in iron dissolution, anaerobic methanogenesis [13], nitrite reduction [14] and anaerobic dinitrogen (N2)–fixation [15]. At any given time, the abundance of CAB would be two to three order less than the seawater, but, if we assume that there is one copepod per litre of seawater, the contribution of CAB to the marine biogeochemical cycles will be significant [1]. Already various studies have shown that CAB has a potential role in biogeochemical processes, such as nitrogen-fixation, [15, 16], denitrification [9], sulphur [17] and iron mineralisation [13].

The masking effect of the abundant bacterial community associated with copepod diet, copepod life stage and environmental conditions was considered the main hindrance in defining core bacterial operational taxonomic units (OTUs; equivalent to species) specific to copepod genera [2, 10]. So, herein, we combined the data from previous studies that dealt with copepod associated bacteria and used machine-learning algorithms to understand the core-bacteria associated with the copepods at least up to the genus level. For this, we analysed 16S rDNA gene sequences (V3-V4 & V4-V5 regions; ~16.5 million reads) of CAB belonging to five different copepod genera (*Acartia* spp., *Calanus* spp., *Centropages* sp., *Pleuromamma* spp. and *Temora* spp.) using Quantitative Insights into Microbial Ecology (QIIME2) package [18]. Also, we hypothesised that, if the copepod genera have specific OTUs, then different copepods have a distinctive CAB, and the biogeochemical potential of the CAB will differ. We used Random Forest classifier, Gradient Boosting Classifier, Principle Coordinate Analysis (PCoA), Analysis Of the Composition of Microbiome (ANCOM), Principle

Component Analysis (PCA) and Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt2) analysis [19] to test this hypothesis. The present study represents one of the biggest CAB-related DNA sequence data analysed to date.

## 2 Materials and Methods

### 2.1 Data collection

We systematically reviewed the studies related to copepod associated bacteriobiome. The relevant published research articles were searched and retrieved from PubMed, Google Scholar and SCOPUS using keywords such as copepods gut microbiome, copepod associated bacteria/microbiome, copepods gut flora, copepod microbiome and zooplankton associated microbiome on Jan 30th, 2020. Apart from the search for the published research articles, we also searched in public databases (for published Ion Torrent, Pyro, and Illumina sequence data) such as the NCBI-SRA, ENA, DDBJ and Figshare using the above-mentioned keywords.

Overall, 11 study data were retrieved for meta-analysis (Table 1) containing 514 next-generation sequence libraries. We have pre-processed separately every individual file within the study and prepared the quality control (QC) report.

### 2.2. Pre-processing

The sequence quality was checked with FastQC tool [20] and the minimum base per quality for future analysis was fixed as PHRED >25. Based on the QC, high rates of erroneous sequences from Illumina, 454 and Ion Torrent files (Table 1) were removed further from the meta-analysis. The two major reasons for the exclusion are 1) erroneous sequences (of PHRED <25) and 2) Short reads (<200 bps) screened by DADA2 [21] while picking sub-Operational Taxonomic Units (s-OTUs). Overall, Illumina sequences contained better quality than the Ion-torrent and Pyrosequence Finally, we carried out a meta-analysis with 452 files of copepods associated bacteriobiome to test the proposed hypothesis.

### 2.3. Meta-analysis

#### 2.3.1. Sequence screening and preparations for meta-analysis

We used Quantitative Insights into Microbial Ecology (QIIME2) version 2019.10 [18], for the meta-analysis. QIIME2 pipeline provides a start-to-finish workflow, beginning with demultiplexing sequence reads and finishing with taxonomic and phylogenetic profiles. The sequences from the individual study were imported to QIIME2 using CasavaOneEight format, and the quality of the sequences was checked by the default settings in QIIME2. Based on the sequence quality, the sequence was trimmed, denoised, aligned and checked for chimera using DADA2 (single and paired-ends sequence were trimmed based on the length of primer used) [21]. The feature table and representative sequence of each file were merged using QIIME2 feature merge table and merged representative sequences.

#### 2.3.2. Taxonomic classification

The merged files were aligned to phylogeny against the Greengenes reference sequence sepp-refs-gg-13-8 using q2-fragment-insertion [22]. Incorrect taxonomic and phylogenetic assignments due to differences in 16S rDNA hypervariable regions and merging the variable

lengths during analysis were solved with q2-fragment insertion technique (SATe-enabled phylogenetic placement in QIIME2 plugin) [22]. The core diversity was calculated before (to calculate the impact on diversity) and after removing mitochondria (mtDNA) and chloroplast (clDNA) sequences from the datasets. The mtDNA and clDNA filtered datasets were further used for calculating diversity, taxonomy, important (core) s-OTUs and the difference in composition estimation using QIIME2 and the diversity graph was plotted within QIIME2. We used Unweighted, Weighted Unifrac and Jaccard distance matrix to compute the beta diversity, and the outcomes were envisaged using Principal Coordinates Analysis (PCoA) in QIIME2. A Permutational Multivariate Analysis of Variance (PERMANOVA) [23] thru the Unweighted, Weighted Unifrac along with Jaccord distance-based beta-diversity was calculated within QIIME2. We used standard pre-trained Greengenes library (gg_13_8_99_OTU_full-length) [24], SILVA reference database (SILVA_188_99_OTUs full-length) [25] and fragment-insertion reference dataset (ref-gg-99-taxonomy). Then we decided to discuss the results from the fragment-insertion reference dataset.

We also implemented the Analysis of the Composition of Microbiome (ANCOM) [26] in QIIME2 plugin to identify the significantly different bacteria between the copepod genera. ANCOM used F-statistics and W-statistics to determine the difference, where W represents the vigour of the ANCOM test for the tested number of species and F represents the measure of the effect size difference for a particular species between the groups (copepods). To predict the important bacteria associated with the copepods, we used sophisticated supervised machine learning classifier (SML): RandomForest Classifier (RFC) [27] and Gradient Boosting Classifier (GBC) [28] using built-in QIIME2. Which is one of the most accurate learning algorithms for managing large and noisy datasets, Random Forest often manages unbalanced sample distributions and is less susceptible to overfitting and generating unbiased classifiers [29]. The gradient boosting method involves the use of several weak learners by taking the loss function from the previous tree and using it to enhance the classification. This technique is less prone to overfitting and does not suffer from the dimensionality curse, but it is susceptible to noisy data and outliers [30].

The mtDNA and clDNA filtered table and representative sequences were also used as an input for predicting CAB potential metabolic function using Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt2) [19]. The output abundance KEGG data were analysed in Statistical Analysis of Taxonomic and Functional Profiles (STAMP) which includes Principle Component Analysis (PCA) [31] to find the significant difference in potential functions of CAB between the copepod genera using Kruskal–Wallis H-test [32] with Tukey–Kramer parameter[33]. The kegg metabolic maps [34-36] was used as a reference to draw the figure representing the copepod genera with a high proportion of potential functional genes.

### 2.4. Copepod phylogeny

The 18S rDNA gene sequences of five copepod genera (used in the present study) were extracted from the Genbank (NCBI). These sequences were aligned and the consensus representative sequence from each genus was obtained using Mega X version. These consensus sequences were used for studying the phylogenetic relationship between the copepods at genera level using Neighbor-joining tree in Mega X [37].

## 3. Results

The present study represented one of the biggest CAB-related DNA sequence data analysed to date. New bioinformatics tools were have been created to cope up with data generated by the next-generation sequencers. To overcome the bias in the tools we used standard, well-recognised pipelines such as FastQC and QIIME2 demultiplexing statistics for reading the quality of sequence, DADA2 algorithm for clustering, aligning, and filtering of chimeric sequences [21]. From the collected data, 12% (n=62, i.e. 35 Roche, six Ion Torrent and 21 Illumina generated sequence files-Table 1) of the files failed during the QC and were omitted and, finally, 452 raw files belonging to five different copepod genera were subjected to analysis.

### 3.1. DNA sequence data analysis

From the 452 raw files, we analysed 16.5 million V3-V4 regions, (except 13 files of V4-V5 archaea specific primer files of Wage et al. [38], Table 1) of bacterial 16S rDNA gene sequences that belong to five copepod genera, i.e. *Acartia* spp., *Calanus* spp., *Centropages* sp., *Pleuromamma* spp. and *Temora* spp. After quality filtering through DADA2 package, an average of 0.1 to 7.8% of sequences was removed (Table 2), and a total of 1, 39, 87,186 sequences were used for downstream analysis.

### 3.2. CAB Alpha diversity

From the bacterial diversity Shannon ('H') indices for the five copepod genera, *Calanus* spp. showed the maximum (Median, (Q1-Q3): 5.85, 4.58-6.29) abundance and evenness of CAB, followed by *Centropages* sp. (5.13, 4.81-5.41) and, the least was observed in *Temora* spp. (2.62, 2.36-2.89) (Figure 1a).

The Kruskal-Wallis analysis revealed that the H index of the CAB within the *Acartia* spp. was significantly different from the CAB of *Calanus* spp., *Centropages* sp., *Temora* spp. and *Pleuromamma* spp. with a p-value ranging between 0.000002 to 0.023779 (Figure 1a). The H index of the CAB within the *Temora* spp. was significantly different from the CAB of *Centropages* sp. (p=0.0012) and *Pleuromamma* spp. (p=0.000209). The H index of the CAB within the *Calanus* spp. was significantly different from the CAB of *Centropages* sp., *Pleuromamma* spp., and *Temora* spp. with a p-value ranging between 0.000008 to 0.05.

Evenness in the indices showed that CAB of the *Calanus* spp. (0.82, 0.67-0.86) have high evenness index followed by *Centropages* sp. (0.74, 0.71-0.77) *Pleuromamma* spp. (0.73, 0.57-0.82), and least in *Temora* spp. (0.65, 0.51-0.68) (Figure 1b).

The Kruskal-Wallis analysis of CAB evenness index was calculated for all the copepod genera (pairwise). There was a significant different evenness (with p-value $\leq 0.05$) between the CAB within *Calanus* spp. and *Acartia* spp., *Pleuromamma* spp., and *Temora* spp. Besides, *Centropages* sp. was significantly different from *Temora* spp. (Figure 1b). The Faith's Phylogenetic Diversity (Faith's_PD) index of the CAB was maximum in the *Pleuromamma* spp. (50.75, 16.41-73.45) and the CAB of *Temora* spp. had less Faith's_PD, (3.59, 2.45-7.26), respectively (Figure 1c).

The variation in the Faith's_PD index of CAB was assessed by Kruskal-Wallis test, which revealed that different copepod genera had highly significant and phylogenetically

distinct bacteriobiome (Figure 1c). Only the CAB within *Acartia* spp. was not significantly different from *Centropages* sp..

### 3.3 CAB Beta diversity

A consensus phylogram of the five copepod genera was constructed (Figure 2a) and compared with the Unweighted UniFrac distance matrix of CAB using PCoA plot. In the present study, from the beta-diversity (PERMANOVA P-value 0.001) patterns, phylogenetically closer *Pleuromamma* spp. and *Calanus* spp. harbour CAB had expressed a mere 7.604 % (axis 1) dissimilarity (Figure 2b); however, the CAB composition still varied between and within copepod genera. As we closely investigate, Unweighted Unifrac distance matrix showed the CAB of *Pleuromamma* spp. and *Calanus* spp. separated into two different clusters (Figure 2b, c), whereas, the CAB of *Calanus* spp. was clustered into a single large cluster in a weighted distance matrix (Figure 2c). In addition to this, in Jaccard distance matrix PCoA revealed that *Calanus* spp. had three distinct CAB clusters (Figure 2d).

On the other hand, the CAB of the phylogenetically closer *Centropages* sp. and *Temora* spp. did show some clustering pattern, but not so distinctive (Figure 2b).

### 3.4. Differential abundance of CAB revealed through ANCOM

ANCOM results showed that a total of 23 CAB phyla, viz., Acidobacteria, Actinobacteria, Bacteroidetes, Chlamydiae, Chlorobi, Crenarchaeota, Cyanobacteria, Elusimicrobia, Euryarchaeota, Firmicutes, Fusobacteria, Gemmatimonadetes, GN02, OD1, [Parvarchaeota], Planctomycetes, Proteobacteria, SBR1093, Spirochaetes, [Thermi], TM6, Verrucomicrobia, and WPS-2 were significantly different between the copepod genera with W and Centred Log-Ratio (clr) statistics ranging between 40 to 30 and 53 to 2.6, respectively (Table S1). The 23-CAB phyla consisted of 32 classes, 78 orders, 145 families and 242 genera which were significantly different between the copepod genera (Supplementary File S1). From these 240 CAB genera, the top two percentile (the W and clr statistical values are given in the supplementary file S1) were chosen to explain the percentile compositional difference of CAB between the copepod genera.

CAB taxa, viz., *Pseudomonas, Anaerospora,* HTCC2207*, Acinetobacter, Ochrobactrum* family Cryomophaceae, Flavobacteriaceae and Methylobacteriaceae (W and clr-statistical values are provided in the supplementary File S1) were found in high percentile within *Calanus* spp. (Figure 3).

Furthermore, from the ANCOM analysis, the CAB taxa, viz., *Paulinella,* RS62*, Candidatus Portiera, Planktotalea, Segetibacter, Octadecabacter,* family Rhodobacteraceae and order Bacteroidales were found in high percentile within *Acartia* spp. (Figure 3). In the case of *Centropages* sp. the CAB genera like *Alteromonas, Pseudoalteromonas, Fluviicola, Oleispira, Ralstonia* and family Colwelliaceae were found to be high percentile. In addition, *Temora* spp. appeared to have high percentile of *Comamonas, Planctomyces, Flavobacterium, Synechococcus, Chryseobacterium* and *Nitrosopumilus*. Only four CAB genera, *Bradyrhizobium, Marinobacter, Photobacterium, and Vibrio,* were significantly high in *Pleuromamma* spp. (Figure 3).

### 3.5. Machine learning-based models to predict important s-OTUs

The overall accuracy of the RandomForest classifier (RFC) model was 0.923 with an accuracy ratio of 1.68, indicating high reliability (Figure 4a). However, the Gradient Boosting Classifier (GBC) model showed better prediction accuracy with a value of 0.967 and accuracy ratio of 1.76 (Figure 4b). The accuracy of RFC in predicting important bacterial s-OTUs in copepod genera was in the range of 0.16 to 1 (Figure 4a), whereas the accuracy of GBC in predicting important s-OTUs in the copepod genera was in the range of 0.16 to 1(Figure 4b). The prediction accuracy of important s-OTUs predicted in *Calanus* spp. and *Pleuromamma* spp. by both of the supervised machine learning (SML) (RFC and GBC) classifiers, was high (1.00) unlike the prediction accuracy for *Acartia* spp. (0.5 in RFC and 0.83 in GBC), *Temora* spp. (0.0 in RFC and 0.66 in GBC) and *Centropages* sp. (0.5 in RFC and 0.5 in GBC). The graphical representation of the machine learning model's Receiver Operating Characteristic (ROC) curve was in the range of 0.98 to 1, and 0.98 to 1 for RFC and GBC, respectively (Figure 4c and 4d). This shows the high positive prediction rate and low false prediction rate for both the SML classifiers (RFC and GBC).

RFC predicted 25 bacterial taxa and one archaeal taxon in five copepod genera as important s-OTUs with differential hierarchical resolutions ranging from family to species level. From the RFC prediction accuracy values, only the s-OTUs predicted as important s-OTUs for the *Calanus* spp. and *Pleuromamma* spp. can be considered due to the low prediction accuracy for *Acartia* spp., *Temora* spp. and *Cetrophages* sp. The following s-OTUs were predicted as the important s-OTUs by RFC only for *Calanus* spp., i.e. *Photobacterium, Vibrio shilonii, Acinetobacter johnsonii, Acinetobacter schindleri, Micrococcus, Micrococcus luteus*, *Anaerospora, and Methylobacteriaceae.* Specific important s-OTUs for the three other genera of copepod was not evident (Figure 4e).

In the case of GBC, a total of 28 taxa and one archaeal taxon was predicted as important s-OTUs for the five copepod genera (Figure 4f). From the GBC prediction accuracy values, only the were s-OTUs predicted as important s-OTUs for the *Calanus* spp. and *Pleuromamma* spp. which can further be considered due to the low prediction accuracy for *Acartia* spp., *Temora* spp. and *Cetrophages* sp. which was similar to the RFC prediction. The following s-OTUs were predicted as the important s-OTUs by GBC only for *Calanus* spp., i.e. *Acinetobacter johnsonii, Vibrio shilonii, Phaeobacter and Piscirickettsiaceae.* And s-OTU of *Marinobacter, Alteromonas, Pseudoalteromonas, Desulfovibrio*, *Limnobacter, Sphingomonas, Methyloversatilis, Enhydrobacter* and Coribacteriaceaewere predicted as important s-OTUs in *Pleuromamma* spp. [58].

### 3.6. Principle Component Analysis reveals that copepod genera do host functionally distinct bacterial diversity.

From the PCA plot on the potential functional genes of CAB, clusters were found for three copepod genera i.e. *Calanus* spp., *Pleuromamma* spp. and *Centropages* sp. (Figure 5).

The potential functional genes of CAB within *Calanus* spp. clustered from the rest of the copepod genera with Principle Component (PC) value of 28.4% in axis 1 and 16.7% in axis 2, whereas the potential functional genes of CAB within *Pleuromamma* spp. showed a variation of 28.4% in axis PC1 and 9.2% in axis PC3. *Centropages* sp. did have unique CAB

functional diversity with a variation of 28.4% in axis PC1 and 9.2% in axis PC3, whereas the potential functional genes of CAB within *Acartia* spp. and *Temora* spp. were scattered.

### 3.7. Biogeochemical potentials of CAB

#### 3.7.1. Potential methanogenesis by CAB: *Evidence of interlinking methanogenesis, DMSP degradation and phosphate utilisation*

The genes responsible for the reduction of methyl phosphonate into methane (MPn genes -phnL, phnM, phnJ, phnI, phnH and phnG) were relatively high in the CAB of *Pleuromamma* spp. and *Calanus* spp. (Figure S2). Also, based on the present analysis, the CAB of the *Centropages* sp. had a high proportion of mttB genes followed by *Acartia* spp. and *Calanus* spp. One should note that these mttB genes involved in the oxidation of Trimethylamine (TMA) to Methyl-CoM (Figure S2).

CAB of *Pleuromamma* spp. and *Calanus* spp. contained some proportion of dmd-tmd (Tri/Demethylation to methylamine) genes, whereas there was no or low proportion of this gene in the CAB of *Temora* spp., *Centropages* sp. and *Acartia* spp. (Figure. S2). The proportion of DmmD gene was high in the CAB of *Centropages* sp. followed by *Acartia* spp. and *Temora* spp., while the CAB of *Pleuromamma* spp. had the minimum proportion. The proportion of dmdA (DMSP to 3-(Methylthio)-propanote) gene was found to be high in the CAB of *Acartia* spp. followed by *Centropages* sp. and *Calanus* spp., whereas, the proportion was minimum in the CAB of *Temora* spp..

Also, the dddL gene (DMSP to Methyl thioether) was found to be high in the CAB of *Centropages* sp. and *Acartia* spp. and low in the CAB of *Temora* spp. But the dmsA gene that converts the DMSO to Methyl thioether was found to be high in the CAB of *Pleuromamma* spp. followed by *Centropages* sp., whereas the genes (dmsB-K00184 and dmsC -K00185) responsible for aerobic conversion of DMSO to Methyl thioether were higher in proportion than the anaerobic genes (dmsA-K07306, dmaB-k07307 and dmsC-K07308), which converts the DMSO to Methyl thioether. The genes dmsB and dmsC (aerobic pathway) were high in CAB of *Temora* spp. followed by *Acartia* spp. whereas, the dmsB and dmsC (anaerobic pathway) gene were high in the CAB of Pleuromamma spp. followed by *Centropages* sp.

Besides, the dmoA gene that converts Methyl thioether to Methanethiol was found only in the *Pleuromamma* spp., but at low proportion. Most importantly, mtsA and mtsB genes (converts Methanethiol to Methyl-CoM) were found to be high in the CAB of *Pleuromamma* spp. when compared to the other copepod genera. Furthermore, the gene responsible for methanogenesis, i.e., the mcrA gene that converts Methyl-CoM to $CH_4$, was found to be high in number within the CAB of *Pleuromamma* spp., but at low proportion (Figure S2).

#### 3.7.2. Methanotrophic potential of CAB

In the present investigation, we found that the relative abundances of mxaF and mxaI genes responsible for methanol dehydrogenases were high in the CAB of *Pleuromamma* spp. with respect to the CAB of other copepod genera (Figure S2). Even though there was a lack of evidence for complete $CH_4$ utilisation, the CAB of *Pleuromamma* spp. had a high number of potential genes responsible for the production of methanol dehydrogenase followed by the CAB of *Centropages* sp. and *Calanus* spp.

### 3.8. Assimilatory sulphate reduction (ASR)

Based on our analysis, in all the copepod genera, ASR pathway genes were predominant rather than the dissimilatory sulphate reduction (DSR) pathway genes. CAB of *Temora* spp. had a higher number of sulfite reductase ferredoxin component (Figure S3a), whereas CAB of *Centropages* sp. had flavoprotein sulfite reductase gene in high proportion (Figure S3b).

### 3.9 Nitrogen fixation

The CAB of the copepod genera was screened for the nifH, nifD and nifK genes responsible for nitrogen fixation. CAB of *Pleuromamma* spp. had a higher proportion of nifH gene followed by *Calanus* spp. whereas *Temora* spp. had a lower proportion (Figure. S4).

### 3.10. Denitrification

Genes involving in all the steps of denitrification (nitrate reductions (narG, napA and napB), nitrite reduction (nirK and nirS), nitric oxide reduction (norB, C) and nitrous-oxide reduction (nosZ) were observed in the CAB of all the five copepod genera, but their relative proportions varied between the genera. The CAB of *Temora* spp. was found to have a high proportion of potential denitrification genes, especially narG, napA and napB genes, followed by the CAB of *Pleuromamma* spp., *Centropages* sp., *Calanus* spp. and *Acartia* spp. (Figure S4).

Among the potential nitrite reductase genes, the proportion of nirK gene was higher than the nirS gene, in all the CAB of copepod genera (Figure S4). Furthermore, the proportion of nirK gene was high in the CAB of *Temora* spp. and *Acartia* spp., whereas a high proportion of nirS was found in *Pleuromamma* spp. and *Calanus* spp. (Figure S4).

The next step in denitrification is the reduction of nitric oxide to nitrous oxide by norB, and norC genes. We also observed from the present analysis, we observed that the CAB of *Temora* spp. have a high proportion of norB gene followed by *Acartia* spp., while the proportion was low in *Pleuromamma* spp. followed by *Calanus* spp. and very low in *Centropages* sp. (Figure S4). Whereas the gene norC was found high in *Pleuromamma* spp. followed by *Calanus* spp. and found to have low in *Temora* spp. (Figure S4). The final reaction is denitrification, i.e. reduction of nitrous oxide to nitrogen by nosZ gene. The CAB of Acartia spp. followed by *Calanus* spp. have a high proportion of nosZ gene (Figure S4).

#### 3.10.1. Anaerobic nitric oxide reduction

The norV (anaerobic nitric oxide reductase) and norW (flavorubredoxin reductase) gene proportion were high in the CAB of *Pleuromamma* spp. compared to the CAB of (descending orders) *Centropages* sp., *Acartia* spp., *Calanus* spp., and least detected in *Temora* spp. (Figure S4).

#### 3.10.2 Dissimilatory nitrate reduction into ammonia (DNRA)

The nrfA gene involves in the final step of DNRA, i.e. reduction of nitrite to ammonia was higher in the CAB of *Calanus* spp., whereas the CAB of *Pleuromamma* spp. and *Centropages* sp. have almost similar proportion (Figure S4).

### 3.11. Carbon processes

Phosphoenolpyruvate Carboxylase (ppc) gene involves in carbon fixation in prokaryotes. This gene was comparatively higher than the other bio-geochemical genes observed in the CAB. While the CAB of *Centropages* sp. have a high proportion of ppc gene, The CAB of *Pleuromamma* spp., *Temora* spp., *Acartia* spp and *Calanus* spp. had the proportion in the descending order (Figure S5a). Also, the CAB of *Centropages* sp. contains have a high proportion of chitinase gene [EC:3.2.1.14] and the least was observed in the CAB of *Calanus* spp. (Figure S5b).

### 3.12. Role of CAB in iron fertilisation

The sequence analysis of CAB showed that the five copepod genera had a different proportion of feoA gene responsible for ferrous iron transport protein A. The CAB of *Temora* spp. have high proportion of feoA gene followed by the CAB of *Pleuromammma* spp., *Acartia* spp. and *Calanus* spp. (Figure S6a), while the other gene fhuF involving in ferric iron reduction was found to be high in the CAB of *Pleuromamma* spp. (Figure S6b).

### 3.13. CAB as a source of cyanocobalamin synthesising prokaryotes

Among the CAB of the five copepod genera analysed, the following was the descending order of copepod genera based on their relative proportion of potential cobalamin synthesising gene: *Temora* spp., *Acartia* spp., *Calanus* spp., *Pleuromamma* spp., and *Centropages* sp. (Figure. S7).

However, from the present study, high proportions of cobalamin synthesising genes in the CAB of *Temora* spp. might be due to the presence of genus *Nitrosopumilus* (phyla *Thaumarchaeota*). We found that the CAB might also be one of the potential sources of cyanocobalamin production in the ocean.

The limitation of the present study could be related to the fact that all the CAB sequences were from the Atlantic Ocean.

## 4. Discussion

### 4.1. CAB diversity between the copepod genera

*Calanus* spp. are filter feeders and mostly herbivores, but they do feed on ciliates and other heterotrophic protists during reproduction and energy shortfall [39, 40]. This might be a reason behind their high H index. Most of the gene sequences used for this meta-analysis were from *Calanus finamarchicus*. However, *Centropages* sp. feeds on different sources from microalgae to fish larvae [41]. *Acartia* spp. are primarily omnivorous (with a high degree of carnivore behaviour), feeding on phytoplankton, occasionally ciliates, and rotifers [42], whereas *Temora* spp. frequently switches its feeding behaviour, i.e. an omnivore or herbivore based on the season and food availability [43]. The bacterial alpha diversity analysis in the *Temora* sp. revealed a significantly lower Shannon diversity. However, in an early study, no difference was reported in the alpha diversity between the *Temora* spp. and *Acartia* spp. [38]. This can be explained based on the source of copepods involved for the study, Wega et al. [38], which was based only on a single source, i.e. the central Baltic sea; however, in our case the CAB sequences for *Acartia* spp. were from the central Baltic sea [38] as well as the Gulf of

Maine [10]. The occurrence of high Faith's_PD in *Pleuromamma* spp. maybe due to their range distribution in the water column and few species within *Pleuromamma* spp. are known to migrate vertically [11, 44] or might also be due to their food uptake, which includes phytoplankton, microzooplankton(ciliates and flagellates) and detritus [11, 45].

The consensus phylogram revealed that, at genera level, *Calanus* spp. was phylogenetically closer to *Pleuromamma* spp. and form two distinct clusters in the PCoA plot. Further, the difference in dissimilarity percentage of CAB between *Pleuromamma* spp. and *Calanus* sp. may be attributed to the difference in vertical migration, life stages and the feeding behaviour between the two copepod genera. *Pleuromamma* spp., an omnivorous feeder [11, 45]. can migrate vertically up to 1000m [11, 44] whereas *Calanus* sp., mostly herbivores, but occasional omnivores [37, 38], can migrate up to 600m [46, 47]. Also, it may be due to the difference in the life stage of *Calanus* (the microbial communities varied between the diapausing and active feeding) [2].

**4.2. ANCOM analysis**

In an early report, bacterial members belonging to the Gammaproteobacteria followed by members of Alphaproteobacteria were observed to be dominant in *Calanus finmarchicus* [10]. But, in the present ANCOM analysis, the high percentile of Gamma and Alphaproteobacteria were equal in number (three genera each) in *Calanus spp.* (Figure 3). Similar to our results, the unclassified genus of Rhodobacteraceae was reported to be abundant in *Acartia longiremis* [10]. Colwelliaceae was reported to be abundant in *Calanus finmarchicus* [10], but, in the present analysis, order Colwelliaceae was found to be in high percentile in *Centropages* sp. An abundance of Flavobacteriaceae was observed along with phytoplankton and diatoms in the gut of *Calanus finmarchicus* with food [2] whereas*Sedinimicola* sp. (Flavobacteriaceae) was observed to be dominant in *Acartia longiremis*., *Calanus finmarchicus* and *Centropages hamatus* [10]. Also, Dorosz et al. [48] reported that *Flavobacterium* was more dominant in *Temora longicornis* than *Acartia tonsa*, whereas, in our case, Flavobacteriaceae was found to be in high percentile in *Calanus* spp.. On comparing, the present ANCOM analysis and previous reports, *Pseudoalteromonas* sp. appeared in high percentile not only within *Centropages* sp. and in *Centropages* sp. [10] but also consistent and abundant bacteria in *Acartia* sp., and *Calanus* sp. [10]. As for as, the prevalence of *Pseudomonas* has been observed in *Pleuromamma* spp. [11], whereas this was not the case in our analysis (Figure 3). Similarly, Cregeen [11] analysed the bacteriobiome of *Pleuromamma* sp. and observed the dominance of *Alteromonas*, but, from our meta-analysis, a high abundance of *Alteromonas* was observed in *Centropages* sp. when compared to five other genera including *Pleuromamma* spp. (Figure 3).

From our analysis, *Nitrosopumilus* was observed to be high in *Temora* spp., but the *Nitrosopumilus* abundance was reported to have no difference between the particle-associated in the water column and within *Temora* spp. [38], so the high percentile observed in our analysis may be due to the exchange of *Nitrosopumilus* from seawater. The Vibrionales was identified as a core member in the gut of *Pleuromamma* spp. [1], which is similar to the present analysis, i.e. *Vibiro* percentile was found to be high in the CAB of *Pleuromamma* spp. The copepods were reported to have a selective niche of *Vibrio* that had capability of degrading chitin [1, 49]. In the present analysis, seven bacterial taxa were found to be in high

percentile in *Centropages* sp. and, among those seven, four taxa belong to the Gammaproteobacteria. A high proportion of Gammaproteobacteria in *Centropages* sp. was also reported earlier [10].

### 4.3. Machine learning-based prediction

The masking effect of the abundant bacterial community associated with the copepod diet and ambient water column should not hinder the detection of core-OTUs, as evidenced by previous studies [1, 2]. QIIME2 core_abundance algorithms used in the present study did not predict single bacterial s-OTUs (data not presented). Hence, we used the machine learning approaches to detect important core sub-OTUs specific to copepod genera.

From our SML classifier results, the important s-OTUs predicted in *Calanus* spp. and *Plueromamma* spp. were found to have high prediction accuracy (AUC=1.00). So, we discuss the important s-OTUs predicted for these two copepod genera (*Calanus* spp. and *Plueromamma* spp.). To begin with, among the important s-OTUs predicted in *Calanus* spp. from the present analysis (both SML models: RFC and GBC) Gammaproteobacteria was a dominant member (14 and 9 s-OTUs from RFC and GBC respectively) followed by Alphaproteobacteria representing six and three s-OTUs from RFC and GBC, respectively. This observation was similar to an earlier study where Gammaproteobacteria and Alphaproteobacteria were reported as core OTUs in *Calanus finmarchicus* [2]. Also, within the Gammaproteobacteria, seven (RFC) and five (GBC) s-OTUs representing the *Acinetobacter* (Moraxellaceae) were predicted as important s-OTUs in the present study. This result was similar to an earlier study in which Moraxellaceae was reported to be closely associated with *Calanus finmarchicus* [10]. Moreover, four s-OTUs of *Acinetobacter* (Moraxellaceae) were also reported as core OTUs in *Calanus finmarchicus* [2]. Adding to it in the present analysis, three s-OTUs from both the SML classifiers RFC and GBC belonging to *Vibrio shilonii* were predicted as important s-OTUs in *Calanus* spp. Comparably, four OTUs of Vibrionaceae (three OTUs of *Vibrio* sp. and one similar to *Vibrio harveyi*) were observed in *Calanus finmarchicus* [2].

In the present SML analysis, one genus, *Bradyrhizobium* (order Rhizobiale), was predicted as an important s-OTUs in *Pleuromamma* spp. by GBC classifiers. Moreover, in the present ANCOM analysis, *Bradyrhizobium* was found to be in high percentile within *Pleuromamma* spp. This *Bradyrhizobium* is also known to have nifH gene, as they usually occur in seawater [50] and also SML-GBC predicted this genus as important s-OTU in *Calanus* spp.. Bradyrhizobiaceae was also found to be the most abundant OTU, i.e.79 of the total 137 sequences in the negative control in a similar analysis [1]. So, in the case of *Bradyrhizobium,* a further investigation shall require to come to a meaningful conclusion.

Moreover, in a previous study, order Vibrionale was also predicted as a core bacterium (based on presence/absence) in *Pleuromamma* spp. [1]. Also, the genus *Pseudoalteromonas* was already reported to occur in high abundance in *Pleuromamma* spp. [11]. However, in the present analysis, GBC predicted five s-OTUs of *Pseudoalteromonas* as important s-OTUs in *Pleuromamma* spp., whereas, the RFC predicted two s-OTUs of *Pseudoalteromonas* as important s-OTUs in *Acartia* spp., *Calanus* spp., and *Centropages* sp. (Figure 4e) This observation was similar to that of *Pseudoalteromonas* reported as a constant and stable OTU

in *Acartia* sp. [38], *Calanus* sp. [2] and *Centropages* sp. [10]. So, it is not wise to consider *Pseudoaltermonas* to be specific to one copepod genera.

In the present study, GBC model predicted three s-OTUs of *Alteromonas* and two s-OTU of *Marinobacter* as important s-OTUs in *Pleuromamma* spp. and the ANCOM analysis also showed that the genus *Marinobacter* proportion was found to be high in *Pleuromamma* spp. Comparably, both the *Alteromonas* and *Marinobacter* were reported to appear commonly in *Pleuromamma* spp. [11]. Even though the abundance of genus *Sphingomonas* was low, it was reported to appear consistently in the *Pleuromamma* spp. [11]. And our analysis predicted this genus as an important s-OTU of *Pleuromamma* spp. (GBC) (Figure 4f).

In the present study, the GBC model predicted *Limnobacter* as an important s-OTU in *Pleuromamma* spp. and the ANCOM analysis also showed that the genus *Limnobacter* proportion was found to be high in *Pleuromamma* spp. Moreover, in a previous study, *Limnobacter* was reported to occur in high abundance as well as unique to copepods (*Pleuromamma* spp.) [11]. Also, the genera Methyloversatilis was reported to be low in abundance in *Pleuromamma* spp., Whereas the present SML -GBC model predicted this genus as an important s-OTU in *Pleuromamma* spp. (Figure 4f). The order Pseuomonadales was reported as a core member in *Pleuromamma* spp. [1]. But our GBC model predicted the bacterial genera *Enhydrobacter* (Pseuomonadales) as an important s-OTU in *Pleuromamma* spp. (Figure 4F). Besides, from the ANCOM analysis, this genus *Enhydobacter* was found to be in high percentile in *Pleuromamma* spp.. But this genus Enhydrobacter was reported to be high in proportion in calanoid copepods [6]. One another important s-OTU predicted in *Pleuromamma* spp. by our GBC model was *Desulfovibrio* and the ANCOM analysis also showed that the genus *Desulfovibrio* proportion was found to be high in *Pleuromamma* spp.

HTCC2207 (Gammaproteobacteria) was predicted as an important s-OTU in *Calanus* spp. by both SML models. Also, from our ANCOM analysis, HTCC2207 was found to be in high percentile in *Calanus* spp. This HTCC2207 is usually more abundant in the seawater and has been reported to be present in a few *Acartia longiremis*., *Calanus finmarchicus*. and *Centropages hamatus* with full gut [10]. Because of their known proteorhodopsin gene and being free water living bacteria [51], the probability of detecting this bacterium in the copepod gut might be due to food ingestion.

*Sediminibacterium* (Chitiniphagaceae) was reported to be in low abundance, but regularly present in *Pleuromamma* spp. [11]. However, in the present analysis RFC model predicted *Sediminibacterium* as important s-OTUs in *Acartia* spp., *Calanus* spp. and *Temora* spp. (Figure 4e and f). Whereas the GBC model predicted *Sediminibacterium* as important s-OTUs in *Acartia* spp. and *Temora* spp.. (Figure 4). Chitiniphagaceae was reported to be associated with calanoid copepods in the North Atlantic Ocean [6]. Earlier studies showed that the genus *Photobacterium* (Phylum: Proteobacteria) was abundant in *Pleuromamma* spp. [11], *Centropages* sp. [11], and *Calanus finmarchicus*. [2]. Herein, *Photobacterium* was detected as an important s-OUT in *Calanus* spp. by the RFC model only. Furthermore, in the present analysis, *Nitrosopumilus* was predicted as an important s-OTU in *Acartia* spp. and *Temora* spp. by both the SML models and this genus, *Nitrosopumilus,* was also reported to be high in percentage in *Acartia* spp. and *Temora* spp. [38].

Further, RFC predicts the Pelomonas as an important s-OTU in *Acartia* spp., *Centropages* sp. and *Calanus* spp. However, in an earlier study, the Pelomonas were ruled out

583 from core OTUs in *Calanus* spp. [2]. Moreover, the GBC predicted two s-OTUs of *RS62* and
584 one s-OTUs of *Planctomyces* as important s-OTUs in *Acartia* spp. and *Temora* spp. This
585 *RS62* belongs to order Burkholderiales, and even though this order was reported to be
586 abundant, their abundance varied between the individual copepods (*Acartia* spp. and *Temora*
587 spp.) [38]. Burkholderiales was also reported as a main copepod associated community [9].
588 However, in the present study, the family Comamonas belonging to Burkholderales was
589 predicted as an important s-OTU in *Acartia* spp., *Temora* spp. by both SML models.
590 About 25 taxa detected by RFC approach were also found to be in high percentile in
591 ANCOM analysis. Among them, eight s-OTUs, i.e. *Anaerospora, Micrococcus, Micrococcus*
592 *luteus, Vibrio shilonii and* Methylobacteriaceae*,* were predicted as important s-OTUs in
593 *Calanus* spp. for the first time in our report (Figure. 4e). From the 28 taxa detected by the
594 GBC model, four s-OTUs, i.e. *Phaeobacter, Acinetobacter johnsonii, Vibrio shilonii,* and
595 Piscirickettsiaceae, were predicted as important s-OTUs in *Calanus* spp. for the first time in
596 our report (Figure 4f). Also, seven s-OTUs i.e. *Marinobacter*, *Limnobacter. Methyloversatilis,*
597 *Desulfovibrio, Enhydrobacter, Sphingobium, Alteromonas and Coriobacteriaceae,* were
598 predicted as important s-OTUs in *Pleuromamma* spp. for the first time in the GBC model.
599
600 **4.4. Potential biogeochemical genes of CAB and their variation and abundance**
601 Bacterial communities exploit copepods as microhabitat by colonising copepods' internal
602 and external surfaces and mediate marine biogeochemical processes [9]. CAB also metabolise
603 organic compounds such as chitin, taurine and other complex molecules in and around the
604 copepod, which could be a hot spot for the biogeochemical process [9]. In one of earlier
605 analysis, potential functional genes in the water column of the Southern Ocean was processed
606 using Parallel-Meta3 software [52], but herein we have used a much advanced PICRUSt2
607 analysis to screen for the potential functional genes.
608 **4.4.1 Methanogenesis**
609 In the present analysis, the bacterial taxa involving in the methane production, *viz*.
610 methanogenesis, methylphosphonate, DMSP and DMSO, were observed in all the copepod
611 genera but the relative proportion varied between them. A similar observation in *Acartia* sp.
612 and *Temora* sp. has been reported [38].
613 In the present analysis, we found that CAB has a complete set of aerobic methanogenesis
614 genes (PhnL, M, J, H and G) which converts methylphosphonate (MPn) to methane ($CH_4$) [53].
615 Some copepods, like *Acartia* sp. and *Temora* sp., were reported to associate with bacteria that
616 involve in $CH_4$ production from MPn [38]. De Corte et al. [9] suggested that different copepod
617 species have different CAB, and only some copepods have specific CAB for methanogenesis
618 and other biogeochemical cycles.
619 A previous study (with 14 C-labelled experiments) observed high methane production in
620 *Temora longicornis* compared to *Acartia spp.* [54]. Also, the methanogenic archeae i.e.
621 *Methanobacterium bryantii-like* sequences and *Methanogenium organophilum,*
622 *Methanolobus vulcani-like sequences* and *Methanogenium organophilum* were noted in
623 *Acartia clausi* and *Temora longicornis* faecal pellets [55]. Meanwhile, in the present study, we
624 observed that *Pleuromamma* spp. has a high proportion of mcrA gene (Figure. S2).
625 *T. longicornis* fed with a high content of TMA/DMA rich phytoplankton produced the
626 maximum amount of $CH_4$, which suggests that this production might be due to the micro-

niches inside the copepods [56]. But in our present analysis, CAB of *Pleuromamma* spp. found to have a high proportion of dmd-tmd gene.

In our meta-analysis *Acartia* spp. found to have a high proportion of dmdA gene. The taxa detected in the present study, like Pelagibacteraceae, some Alpha and Gammaproteobacteria, are known to have dmdA genes [57].

Copepods feeding on phytoplankton liberate DMSP, which, in turn, is utilised by the DMSP-consuming bacteria in the gut (*Acartia tonsa*), leading to methane production [58]. Moreover, the methane enrichment in the Central Baltic Sea was due to the dominant zooplankton *Temora longicornis* feeding on the DMSP/DMSO-rich Dinophyceae resulting in methane release [54].

Instead of analysing faecal pellets [58] and anaerobic incubation experiments [59], further research should also consider CAB-mediated aerobic methanogenesis as one of the factors to solve the 'Ocean methane paradox'.

#### 4.4.1.1. Methanotrophic potential of CAB

The present analysis showed that the CAB of *Pleuromamma* spp. and *Calanus* spp. were found to have a high proportion of methanol dehydrogenase genes (mxaF and mxaI) (Figure S2). This may be due to the presence of Proteobacteria that involves methane oxidation viz. Beijerinckiaceae, Methylococcaceae Methylocystaceae and Methylacidiphilaceae (Verrucomicrobia) (Supplementary File-S2) [60].

### 4.4.2. Assimilatory sulphate reduction (ASR)

A relative abundance of genera like *Synechococcus* and the Deltaproteobacterial family (unclassified genera in Desulfovibrionaceae), Rhodobacteriaceae and Flavobacterium (Supplementary File S2) was observed in CAB of *Temora* spp. and this might be responsible for the ASR pathway, as these genera are known to have ferredoxin-sulfite reductase activity (Supplementary File-S2).

### 4.4.3. Nitrogen fixation

The abundance of nifH gene was reported to be high in copepods collected from the coastal waters of Denmark (Øresund) (mostly contributed by *Acartia* spp.) with *Vibrio* spp. as dominant members [16]. But, in the present study, the nifH gene was found to be high in the CAB of *Pleuromamma* spp (Figure S4). and one should note that this may be due to the high abundance of genus *Vibrio* in the CAB of *Pleuromamma* spp. (Supplementary File -S2). *Vibrio* attached to the exoskeleton, and gut lining of copepods [61] use chitin as carbon and energy source was reported in the past [10]. Furthermore, copepods are reported to be a hotspot for nitrogen fixation at a rate of 12.9–71.9 μmol N $dm^{-3}$ copepod biomass per day [16]. The abundance of nifH gene in the CAB of *Pleuromamma* spp. maybe due to the presence of genera like *Synechococcus, Prochlorococcus, Bradyrhizobium, Microcystis,* and *Trichodesmium* (Supplementary File S2).

#### 4.4.3.1. Denitrification

In our analysis, the CAB of *Temora* spp. followed by *Pleuromamma* spp. were found to have a high proportion of napA and napB genes (Figure S4), whereas the abundance of napA and narG genes were reported in the North Atlantic copepods contributed by *Calanus* sp. and *Paraeuchaeate* sp. [9]. However, in the present analysis, the CAB of *Temora* spp. was found to have a high proportion of narG (Figure S4). Bacterial genera like *Pseudoalteromonas Actinobacterium* and *Shewanella,* also the nirS gene were reported in live and dead *Calanus finmarchicus* [14]. Likewise, from our analysis, both *Pseudoalteromonas* and *Actinobacterium* were found in *Calanus* spp. Also, metagenome analysis of copepod associated microbial community was reported to have genes responsible for denitrification and DNRA [9].

**4.4.3.2. Anaerobic nitric oxide reduction**

Families like Aeromonadaceae and Enterobacteriaceae were observed in the CAB of *Pleuromamma* spp. and *Calanus* spp. (in relatively higher proportion than other copepods). The genera *Aeromonas* (family Aeromonadaceae) [62] and *Escherichia coli* (family Enterobacteraceae) [63] are known to have norV genes. The presence of these bacterial taxa in *Pleuromamma* spp. maybe due to feeding of ciliates, flagellates and detritus particles [11, 45]. This may be one of the reasons to have a high proportion of norV and norW genes in these copepods (Figure S4).

**4.4.4. Carbon processes**

Bacterial genera, like some genera in Colwelliaceae [10, 64] *Flavobacterium, Arthrobacter, Serratia, Bacillus, Enterobacter, Vibrio,* [65] *Pseudoalteromonas* [64] and *Achromobacter* [66] produce chitinase. The presence of chitinase gene in CAB is not surprising as their foregut and hindgut are made up of chitin [11]. The overall outline of CAB-mediated biogeochemical pathways is represented in Figure 6.

**4.4.5. Role of CAB in iron fertilisation**

*Pleuromamma* spp. carries a similar proportion of ferric iron reductase gene (fhuF) and ferrous iron transport protein A gene (feoA) (Figure S6a, b). The presence of high proportion of ferric iron reductase gene fhuF in *Pleuromamma* spp. needs detailed investigation. It was reported that acidic and low-oxygen conditions in copepod gut may assist iron dissolution and remineralisation, forming soluble Fe (II) [13, 67]. Thus, this increases the iron bioavailability in the surrounding, which promotes phytoplankton growth [67]. Also, bacterial community associated with the zooplankton, such as Bacteroidetes, Alphaproteobacteria and Gammaproteobacteria, are known to carry genes involved in iron metabolism [9].

An early study on *Thalassiosira pseudonana* fed to *Acartia tonsa* was found to have Fe in the faecal pellets [68]. But, in the present analysis, *Acartia* spp. was found to have less proportion of feoA gene when compared to *Temora* spp. and *Pleuromamma* spp. Moreover, genes involved in iron metabolism were reported to be high in zooplankton associated microbiome [9].

Since the differential iron contributions of different copepod genera were unknown until now. For organisms that must combat oxygen limitation for their survival

(*Pleuromamma* spp.), pathways for the uptake of ferrous iron are essential. Nevertheless, the meta-analysis performed here showed that *Pleuromamma* spp. could be a significant contributor to both iron bioavailability and nitrogen fixation.

#### 4.4.6. CAB as a source of cyanocobalamin synthesising prokaryotes

Organisms within all domains of life require the cofactor cobalamin (vitamin B12), which is usually produced only by a subset of bacteria and archaea [69]. Previous studies reported that the cobalamin in ocean surface water is due to de nova synthesis by Thaumarchaeota. Moreover, few members of Alphaproteobacteria, Gammaproteobacteria and Bacteroidetes genomes were reported to have the cobalamin synthesising gene [69]. In our analysis, the CAB of *Temora* spp. was found to have a high proportion of Thaumarchaeota, whereas Alpha-gammaproteobacteria were found to be high in the CAB of *Acartia* spp., *Calanus* spp. and *Pleuromamma* spp. In this regard, further studies on CAB diversity from different ocean realms would shine a light on the actual potential of CAB in the global biogeochemical cycles.

## 5. Conclusion

Herein, five copepod genera viz. *Acartia* spp., *Calanus* spp., *Centropages* sp., *Pleuromamma* spp., and *Temora* spp., and their associated bacteriobiome were investigated. The use of meta-analysis in the present study reveals the difference in bacterial diversity indices within the alpha and beta-diversity. To be more specific, the meta-analysis showed significant variations in the alpha diversity between the copepod genera. Moreover, it revealed that *Calanus* spp have high Shannon index (H-index) and *Pleuromamma* spp. have high Faith's Phylogenetic Diversity. Furthermore, the meta-analysis revealed that the CAB within the phylogenetically closer *Pleuromamma* spp. and *Calanus* spp. expressed a mere 7.604% (axis 1) dissimilarity distance in PCoA analysis (Unweighted Unifrac distance matrix based on the phylogenetic index). Likewise, from the meta-analysis, we were able to identify the bacterial taxa which are significantly abundant in each copepod genera in comparison with others.

In earlier studies, the core bacterial OTUs were identified based on their presence/absence [1] as well as by using distribution-based clustering (DBC) algorithms [2]. Herein, machine learning models were used to predict the important copepod associated bacterial genera within the five different copepod genera. In specific, we used supervised machine learning models to predict the important bacterial s-OTUs. We predicted 28 bacterial taxa and one archeael taxon (SML-GBC) as important s-OTUs in the five copepod genera. Among the predicted bacterial genera, in common, *Vibrio shilonii, Acinetobacter johnsonii, Piscirickeesiaceae,* and *Phaeobacter* were reported as important s-OTUs in the *Calanus* spp. and *Marinobacter, Limnobacter. Methyloversatilis, Desulfovibrio, Enhydrobacter, Sphingobium, Alteromonas and Coriobacteriaceae* were predicted as important s-OTUs in *Pleuromamma* spp. for the first time. Additionally, the prediction accuracy (for *Calanus* spp. and *Pleuromamma* spp.) of the machine learning models used here showed high accuracy, which indicates the reliability of the predicted important s-OTUs in the copepod genera. Notably, from the machine learning-based classification it was evident that specific bacterial s-OTUs do exist for copepods.

Furthermore, our meta-analysis revealed that the five copepod genera have bacterial communities that are capable of mediating methanogenesis (with evidence of interlinking the methane production, DMSP degradation and phosphate utilisation) and methane oxidation. We also found the five copepod genera to have more potential Assimilatory Sulfur Reducing (ASR) microbial communities than the Dissimilatory Sulfate Reducing (DSR) communities within the CAB. Likewise, the bacterial community with potential genes involved in nitrogen fixation, denitrification and DNRA were also observed among the CAB of these five copepod genera. We also found the potential genes that perform carbon fixation, iron remineralisation and Cyanocobalamin (vitamin B12) synthesis in the CAB of the five copepod genera.

**Authors Contribution**

The authors' BS, MS, PC and MG designed the work., BS executed out the meta-analysis and machine learning (QIIME) approach., PC helped in constructing the copepod phylogenetic tree. UN and PC helped in data arrangement and review of the literature. MS helped in executing machine learning approach. BS, MS, and PC wrote the initial draft. Editing and rewriting were performed by MS and MG.

**Acknowledgements**

The authors thank the Director, CSIR-NIO, for encouraging this work. BS, PC, UVN and MG received the financial assistance from the Council of Scientific & Industrial Research, Government of India, under projects OLP2005 & MLP1802. MS is also funded by the Engineering and Physical Sciences Research Council, UK, and Imperial College London (EP/N509486/1: 1,979,819). We thank our funders. We also thank the High-performance computing facility "Pravah" to carry out the bioinformatics and Machine Learning work. This is NIO's contribution No___. The authors declare that they have no conflict of interest. All the copepod associated microbiome sequence datasets were downloaded from the NCBI SRA database. The information on their NCBI BioProject numbers (accession numbers), species names, 16S rDNA regions, sequencing platforms used along with the corresponding references can be found in the supplementary file Table 1. All of these data are free to download from the NCBI database (https://www.ncbi.nlm.nih.gov/sra). In the next step, we vetted these crude files for quality reads. In this study, only this quality reads were used for our meta-analysis. These datasets in both sequence format, as well as biom format, is submitted in the Figshare database and is free to download from the following link (https://doi.org/10.6084/m9.figshare.c.5086811.v1). This dataset was used for taxonomy, machine-learning, Picrust2 and ANCOM analysis. The dataset enclosing the results for the analyses mentioned above was also deposited in the Figshare repository and is free to download from the following link (https://doi.org/10.6084/m9.figshare.c.5087183.v3). Anyone can download and reuse this data for their analysis.

**References**

1 Shoemaker and Moisander, 2017 Shoemaker, K. M. & Moisander, P. H. Microbial diversity associated with copepods in the North Atlantic subtropical gyre. FEMS

Microbiology Ecology 91, (2015). https://doi.org/10.1093/femsec/fiv064. Assessed on 15-01-2020. Reproduced from NCBI/SRA (PRJNA248671).

2 Datta, M. S. et al. Inter-individual variability in copepod microbiomes reveals bacterial networks linked to host physiology. ISME J 12, 2103–2113 (20 https://doi.org/10.1038/s41396-018-0182-1. Data Assessed on 15-01-2020. Reproduced from NCBI/SRA (PRJNA322089).

3 Steinberg, D. K. et al. Zooplankton vertical migration and the active transport of dissolved organic and inorganic carbon in the Sargasso Sea. Deep Sea Research Part I: Oceanographic Research Papers 47, 137–158 (2000).

4 Chen, M., Kim, D., Liu, H. & Kang, C.-K. Variability in copepod trophic levels and feeding selectivity based on stable isotope analysis in Gwangyang Bay of the southern coast of the Korean Peninsula. Biogeosciences 15, 2055–2073 (2018).

5 Tang, K. Copepods as microbial hotspots in the ocean: effects of host feeding activities on attached bacteria. Aquat. Microb. Ecol. 38, 31–40 (2005).

6 De Corte, D. et al. Linkage between copepods and bacteria in the North Atlantic Ocean. Aquat. Microb. Ecol. 72, 215–225 (2014).

7 Grossart HP, Dziallas C, Leunert F, Tang KW. Bacteria dispersal by hitchhiking on zooplankton. Proc Natl Acad Sci USA 107: 11959−11964 (2010).

8 Tang, K., Turk, V. & Grossart, H. Linkage between crustacean zooplankton and aquatic bacteria. Aquat. Microb. Ecol. 61, 261–277 (2010).

9 De Corte, D. et al. Metagenomic insights into zooplankton-associated bacterial communities. Environ Microbiol 20, 492–505 (2017).

10 Moisander, P. H. et al. Copepod-Associated Gammaproteobacteria Respire Nitrate in the Open Ocean Surface Layers. Front. Microbiol. 9, (2018).

11 Cregeen, S.J.J. . Microbiota of dominant Atlantic copepods: *Pleuromamma* sp. as a host to a betaproteobacterial symbiont. Ph.D., Thesis, University of Southampton, pp-1-183.(2016).

12 Marchesi, J. R. & Ravel, J. The vocabulary of microbiome research: a proposal. Microbiome 3, (2015).

13 Tang, K. W., Glud, R. N., Glud, A., Rysgaard, S. & Nielsen, T. G. Copepod guts as biogeochemical hotspots in the sea: Evidence from microelectrode profiling of *Calanus* spp. Limnol. Oceanogr. 56, 666–672 (2011).

14 Glud, R. N. et al. Copepod carcasses as microbial hot spots for pelagic denitrification. Limnol. Oceanogr. 60, 2026–2036 (2015).

15 Proctor, L. Nitrogen-fixing, photosynthetic, anaerobic bacteria associated with pelagic copepods. Aquat. Microb. Ecol. 12, 105–113 (1997).

16 Scavotto, R. E., Dziallas, C., Bentzon-Tilia, M., Riemann, L. & Moisander, P. H. Nitrogen-fixing bacteria associated with copepods in coastal waters of the North Atlantic Ocean. Environ Microbiol 17, 3754–3765 (2015).

17 Dong, Y., Yang, G.-P. & Tang, K. W. Dietary effects on abundance and carbon utilization ability of DMSP-consuming bacteria associated with the copepod *Acartia tonsa* Dana. Marine Biology Research 9, 809–814 (2013).

18 Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37, 852–857 (2019).
19 Douglas, G. M. et al. PICRUSt2 for prediction of metagenome functions. Nat Biotechnol 38, 685–688 (2020).
20 Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37, 852–857 (2019).
21 Callahan, B. J. et al. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 13, 581–583 (2016).
22 Janssen, S. et al. Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information. mSystems 3, (2018).
23 Anderson, M. J. Permutational Multivariate Analysis of Variance (PERMANOVA). Wiley StatsRef: Statistics Reference Online 1–15 (2017) doi:10.1002/9781118445112.stat07841.
24 McDonald, D. et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 6, 610–618 (2011).
25 Bokulich, N. et al. bokulich-lab/RESCRIPt: 2020.11. (Zenodo, 2020). doi:10.5281/ZENODO.3891931.
26 Mandal, S. et al. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microbial Ecology in Health & Disease 26, (2015).
27 Breiman, L. Machine Learning 45, 5–32 (2001).
28 Friedman, J. H. machine. Ann. Statist. 29, 1189–1232 (2001).
29 Roguet, A., Eren, A. M., Newton, R. J. & McLellan, S. L. Fecal source identification using random forest. Microbiome 6, (2018).
30 Dhoble, A. S., Lahiri, P. & Bhalerao, K. D. Machine learning analysis of microbial flow cytometry data from nanoparticles, antibiotics and carbon sources perturbed anaerobic microbiomes. J Biol Eng 12, (2018).
31 Parks, D. H., Tyson, G. W., Hugenholtz, P. & Beiko, R. G. STAMP: statistical analysis of taxonomic and functional profiles. Bioinformatics 30, 3123–3124 (2014).
32 Kruskal, W. H. & Wallis, W. A. Use of Ranks in One-Criterion Variance Analysis. Journal of the American Statistical Association 47, 583–621 (1952).
33 Tukey–Kramer Method. in Encyclopedia of Systems Biology 2304–2304 (Springer New York, 2013). doi:10.1007/978-1-4419-9863-7_101575.
34 Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28, 27–30 (2000).
35 Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. Protein Science 28, 1947–1951 (2019).
36 Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Research (2020) doi:10.1093/nar/gkaa970.
37 Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular Biology and Evolution 35, 1547–1549 (2018).

884 38 Wäge, J. et al. Microcapillary sampling of Baltic Sea copepod gut microbiomes
885 indicates high variability among individuals and the potential for methane production.
886 FEMS Microbiology Ecology 95, (2019).
887 39 Ohman, M. D. & Runge, J. A. Sustained fecundity when phytoplankton resources are
888 in short supply: Omnivory by Calanus finmarchicus in the Gulf of St. Lawrence.
889 Limnol. Oceanogr. 39, 21–36 (1994).
890 40 Harris, R. Feeding, growth, and reproduction in the genus *Calanus*. ICES Journal of
891 Marine Science 57, 1708–1726 (2000).
892 41 Saage, A., Vadstein, O. & Sommer, U. Feeding behaviour of adult *Centropages*
893 *hamatus* (Copepoda, Calanoida): Functional response and selective feeding
894 experiments. Journal of Sea Research 62, 16–21 (2009).
895 42 Chen, M., Liu, H. & Chen, B. Seasonal Variability of Mesozooplankton Feeding
896 Rates on Phytoplankton in Subtropical Coastal and Estuarine Waters. Front. Mar. Sci.
897 4, (2017).
898 43 Dam, H. G. & Lopes, R. M. Omnivory in the calanoid copepod *Temora longicornis*:
899 feeding, egg production and egg hatching rates. Journal of Experimental Marine
900 Biology and Ecology 292, 119–137 (2003).
901 44 Buskey, E.J., Baker, K.S., Smith, R.C. & Swift, E. Photosensitivity of the oceanic
902 copepods *Pleuromamma gracilis* and *Pleuromamma xiphias* and its relationship to
903 light penetration and daytime depth distribution. Marine Ecology Progress Series.
904 55:207–216 (1989).
905 45 Wilson, S. & Steinberg, D. Autotrophic picoplankton in mesozooplankton guts:
906 evidence of aggregate feeding in the mesopelagic zone and export of small
907 phytoplankton. Mar. Ecol. Prog. Ser. 412, 11–27 (2010).
908 46 Hirche, H. J. Overwintering of *Calanus finmarchicus* and *C. helgolandicus*. Mar.
909 Ecol. Prog. Ser. *11*, 281–290 (1983).
910 47 Tande, K. S. An evaluation of factors affecting vertical distribution among recruits of
911 *Calanus finmarchicus* in three adjacent high-latitude localities. in Biology of
912 Copepods 115–126 (Springer Netherlands, 1988). doi:10.1007/978-94-009-3103-
913 9_10.
914 48 Dorosz, J., Castro-Mejia, J., Hansen, L., Nielsen, D. & Skovgaard, A. Different
915 microbiomes associated with the copepods *Acartia tonsa* and *Temora longicornis*
916 from the same marine environment. Aquat. Microb. Ecol. 78, 1–9 (2016).
917 49 Heidelberg, J. F., Heidelberg, K. B. & Colwell, R. R. Bacteria of the γ-Subclass
918 Proteobacteria Associated with Zooplankton in Chesapeake Bay. AEM 68, 5498–5507
919 (2002).
920 50 Jayakumar, A. & Ward, B. B. Diversity and distribution of Nitrogen Fixation Genes in
921 the Oxygen Minimum Zones of the World Oceans. (2020) doi:10.5194/bg-2019-445.
922 51 Stingl, U., Desiderio, R. A., Cho, J.-C., Vergin, K. L. & Giovannoni, S. J. The SAR92
923 Clade: an Abundant Coastal Clade of Culturable Marine Bacteria Possessing
924 Proteorhodopsin. AEM 73, 2290–2296 (2007).
925 52 Sadaiappan, B. et al. Metagenomic 16S rDNA amplicon data of microbial diversity
926 and its predicted metabolic functions in the Southern Ocean (Antarctic). Data in Brief
927 28, 104876 (2020).

928 53 Yao, M., Henny, C. & Maresca, J. A. Freshwater Bacteria Release Methane as a By-
929 Product of Phosphorus Acquisition. Appl. Environ. Microbiol. 82, 6994–7003 (2016).
930 54 Stawiarski, B. et al. Controls on zooplankton methane production in the central Baltic
931 Sea. Biogeosciences 16, 1–16 (2019).
932 55 Ditchfield, A. et al. Identification of putative methylotrophic and hydrogenotrophic
933 methanogens within sedimenting material and copepod faecal pellets. Aquat. Microb.
934 Ecol. 67, 151–160 (2012).
935 56 de Angelis, M. A. & Lee, C. Methane production during zooplankton grazing on
936 marine phytoplankton. Limnol. Oceanogr. 39, 1298–1308 (1994).
937 57 Howard, E. C., Sun, S., Biers, E. J. & Moran, M. A. Abundant and diverse bacteria
938 involved in DMSP degradation in marine surface waters. Environmental
939 Microbiology 10, 2397–2410 (2008).
940 58 Tang, K. W., Visscher, P. T. & Dam, H. G. DMSP-consuming bacteria associated with
941 the calanoid copepod *Acartia tonsa* (Dana). Journal of Experimental Marine Biology
942 and Ecology 256, 185–198 (2001).
943 59 Ploug, H., Kühl, M., Buchholz-Cleven, B. & Jørgensen, B. Anoxic aggregates - an
944 ephemeral phenomenon in the pelagic environment? Aquat. Microb. Ecol. 13, 285–
945 294 (1997).
946 60 Tamas, I., Smirnova, A. V., He, Z. & Dunfield, P. F. The (d)evolution of
947 methanotrophy in the Beijerinckiaceae—a comparative genomics analysis. ISME J 8,
948 369–382 (2013).
949 61 Rawlings, T. K., Ruiz, G. M. & Colwell, R. R. Association of Vibrio cholerae O1 El
950 Tor and O139 Bengal with the Copepods *Acartia tonsa* and Eurytemora affinis. AEM
951 73, 7926–7933 (2007).
952 62 Liu, J. et al. Diverse effects of nitric oxide reductase NorV on *Aeromonas hydrophila*
953 virulence-associated traits under aerobic and anaerobic conditions. Vet Res 50, (2019).
954 63 Gardette, M., Daniel, J., Loukiadis, E. & Jubelin, G. Role of the Nitric Oxide
955 Reductase NorVW in the Survival and Virulence of Enterohaemorrhagic *Escherichia*
956 *coli* during Infection. Pathogens 9, 683 (2020).
957 64 Cottrell, M. T., Wood, D. N., Yu, L. & Kirchman, D. L. Selected Chitinase Genes in
958 Cultured and Uncultured Marine Bacteria in the α- and γ-Subclasses of the
959 Proteobacteria. Appl. Environ. Microbiol. 66, 1195–1201 (2000).
960 65 Donderski, W., & Trzebiatowska, M. Influence of physical and chemical factors on
961 the activity of chitinases produced by planktonic bacteria isolated from Jeziorak Lake.
962 Polish Journal of Environmental Studies, 9(2), 77–82 (2000).
963 66 Subramanian, K. et al. Bioconversion of chitin and concomitant production of
964 chitinase and N-acetylglucosamine by novel *Achromobacter xylosoxidans* isolated
965 from shrimp waste disposal area. Sci Rep 10, (2020).
966 67 Schmidt, K. et al. Zooplankton Gut Passage Mobilizes Lithogenic Iron for Ocean
967 Productivity. Current Biology 26, 2667–2673 (2016).
968 68 Hutchins, D. A., Wang, W.-X. & Fisher, N. S. Copepod grazing and the
969 biogeochemical fate of diatom iron. Limnol. Oceanogr. 40, 989–994 (1995).

69 Doxey, A. C., Kurtz, D. A., Lynch, M. D., Sauder, L. A. & Neufeld, J. D. Aquatic metagenomes implicate Thaumarchaeota in global cobalamin production. ISME J 9, 461–471 (2014).
70 Skovgaard, A., Castro-Mejia, J. L., Hansen, L. H. & Nielsen, D. S. Host-Specific and pH-Dependent Microbiomes of Copepods in an Extensive Rearing System. PLoS ONE 10, e0132516 (2015).
71 Shoemaker, K. M. & Moisander, P. H. Microbial diversity associated with copepods in the North Atlantic subtropical gyre. FEMS Microbiology Ecology 91, (2015).
72 Shelyakin, P. V. et al. Microbiomes of gall-inducing copepod crustaceans from the corals Stylophora pistillata (Scleractinia) and Gorgonia ventalina (Alcyonacea). Sci Rep 8, (2018).

**Table 1.** List of sequence libraries representing the copepods-associated bacteribiome. Out of these, only seven libraries (highlighted in red font) were analysed in this study.

**Table 2.** Details of the number of Illumina files, sequences extracted, and quality filtered (Phred score <25). RP indicate 'relative proportion'.

**Figure 1.** Alpha diversity composition and variation a) Shannon index (Richness and diversity accounting for both abundance and evenness of the taxa present); b) Evenness index (Relative evenness of species richness); c) Faith's Phylogenetic Diversity index (biodiversity incorporating phylogenetic difference between species) corresponding to the CAB within five different copepod genera.

**Figure 2.** a) Unweighted Unifrac distance matrix (community dissimilarity that incorporates phylogenetic relationships between the features); b) Weighted Unifrac distance matrix (community dissimilarity that incorporates phylogenetic relationships between the features); c) Jaccord distance-based beta-diversity. The CAB of representative copepods are colour coded; d)18S rDNA phylogenetic tree of five copepod genera used in the study.

**Figure 3.** Top two percentile of the CAB-bacterial genera observed in the copepods obtained via ANCOM.

**Figure 4.** a) Confusion matrix for the RandomForest Classifier (RFC) model; b) Confusion matrix for the Gradient Boosting Classifier (GBC) model; c) Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) for the RFC model; d) Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) for the GBC model; e) Heatmap of the predicted important s-OTUs in the five copepod genera using RFC; f) Heatmap of the predicted important s-OTUs in the five copepod genera using GBC.

**Figure 5.** PCA plot for overall diversity pattern of potential functional genes observed among the CAB within the five copepod genera.

**Figure 6.** Overall representation of the potential functional genes of CAB involved in biogeochemical cycles. The circle and the colour represent the copepod genera contained in high proportion for that particular biogeochemical process.