PNAS
www.pnas.org

Supplementary Information for:

Microbial Dynamics of Elevated Carbon Flux in the Open Ocean's Abyss

Kirsten Poff, Andy O. Leu, John M. Eppley, David M. Karl, and Edward F. DeLong

Corresponding author: Edward F. DeLong
Email: edelong@hawaii.edu

This PDF file includes:

Supplementary Information Text

SI Appendix Figures S1 to S9

SI Appendix Table S1 to S7

Captions for Datasets S1 to S9

References cited

Other supplementary materials for this manuscript include the following:

Datasets S1 to S9

Supplementary Information Text

Detailed methods

Sample collection

Sinking POM was collected and preserved in sequencing sediment traps deployed at 4000 m over a three-year time period as previously described (1). Moored sequential sediment traps were deployed at 4,000 m at station ALOHA (McLane Research Laboratories Inc.). These traps were deployed yearly and operated between periods of 8-9 months during 2014, 2015 and 2016. For each year, 21 samples were collected for a total of 63 samples collected over the entire time-series. Each sample was collected over a 10-12-day interval (see Dataset S1 for details). Sinking particles captured in the trap were preserved in a nucleic acid preservation solution as previously described (1). A second trap was deployed over the same time period collected sinking particles in a formalin brine solution for analysis of particulate carbon and nitrogen as described in Karl et al. (2). After sediment trap recovery and deployment, all samples were processed, and DNA and RNA extracted and purified, exactly as described in Boeuf et al. (1).

Sample processing and data generation

Metatranscriptomic libraries were prepared using a ScriptSeq v2 RNA-seq kit (Illumina, San Diego, CA). Unique barcodes were added to the cDNA by PCR enrichment. Sequencing was performed on an Illumina NextSeq 500 system using a V2 high output 300 cycle kit (Illumina, San Diego, CA). Primer sequences were removed from sequence reads using Trimmomatic 0.38 (parameters: ILLUMINACLIP:2:30:10) (3); read pairs were assembled using FLASH 1.2.11 (default parameters) (4) and low quality bases were removed with Trimmomatic (parameters: LEADING:10 TRAILING:10 MINLEN:100). Reads containing spiked in sequences were removed using the bbduk script from BBMap 38.22 (5). The first pass removed

phiX sequence and reads with extreme GC values (parameters: k=27 hdist=1 qtrim=rl trimq=17 cardinality=t mingc=0.05 maxgc=0). The second pass removed and counted the quantitative standards (parameters: k=27 hdist=1 cardinality=t).

Metagenomic libraries were prepared using an automated NeoPrep instrument with TruSeq Nano DNA library preparation kit (Illumina #NP1011001, San Diego, CA), using an input of 25 ng of sheared genomic DNA. Libraries were sequenced either using a 150 bp paired-end NextSeq500/550 High Output v2 reagent kit (Illumina #FC4042004, San Diego, CA).

Quality controlled sequences were compared to all ribosomal RNA (rRNA) models from RFAM release 12.1 (6) using cmsearch (parameters: --hmmonly) from infernal 1.1.2 (7). A total of 662 million reads with rRNA sequences passed the quality control filters. The taxonomic affiliations of the SSU rRNA sequences were assessed by homology to the SILVA SSU NR99 database Release 132 (8) using BWA 0.7.15-r1140 (9) with matches limited to at least 97% identity over at least 70 bases.

Diversity Calculations for SSU rRNA taxon determinations

Family level Eukaryote and combined Bacteria and Archaea datasets were rarefied to their respective minimum depths and Shannon diversity indices were calculated for each sample using the R package 'vegan' (10). Using the R package 'stats', a dependent 2-group Wilcoxon Signed Rank Test was performed (11).

Recovery of the MAGs

A total of 3.19 billion metagenomic raw paired-end reads were quality controlled and assembled as follows. First, contaminant sequences were removed using the bbduk script from

BBMap 38.22 (5) in two passes. The first pass used parameters “ktrim=r k=23 mink=11 hdist=1 tbo tpe tbo tpe” for Illumina sequencing adapters, and the second pass used parameters “k=27 hdist=1 qtrim=rl trimq=17 cardinality=t mingc=0.05 maxgc=0.95” to remove phiX, low quality bases, and sequences with unrealistically high or low GC. Next, additional low-quality bases and sequences were removed with Trimmomatic 0.38 (parameters: LEADING:10 TRAILING:10 MINLEN:100) (3). Unpaired reads were removed using the dropse command from seqtk 1.2 (12), producing a total of 2.45 billion quality-controlled metagenomic reads across 63 samples. Finally, the cleaned reads from each sample were assembled using SPAdes 3.11.1 (parameters: -meta -k 21,33,55,77,99,127) (13) generating 92 million contigs, 900,000 of these were longer than 1500bp.

Mapping of quality reads was performed using CoverM v0.3.1 with default parameters (https://github.com/wwood/CoverM). For each assembled metagenome, metagenomic assembled genomes were recovered using MetaBAT1 v0.32.5 (14) with all the sensitivity settings, MetaBAT2 v2.12.1 (15) and MaxBin2 v2.2.6 (16) using the 40 and 107 gene sets. The resulting MAGs from each assembly were assessed for completeness and dereplicated using DAStool (17). Completeness and contamination rates of the MAGs were assessed using CheckM v1.0.13 (18) with the ‘lineage wf’ command. MAGs from all metagenomes were subsequently dereplicated using dRep v2.2.3 (19) using the dereplicate wf at ≥97% average nucleotide identity over ≥70% alignment and the best MAGs were chosen based on genome completeness. The dereplicated MAGs were further refined by reassembling the mapped quality trimmed reads with SPAdes (20) using the –careful and –trusted-contigs setting. Additional scaffolding and resolving

ambiguous bases of the MAGs was performed using the 'roundup' mode of FinishM v0.0.7 (https://github.com/wwood/finishm).

Taxonomic inference of the MAGs

Classification of the MAGs was determined using GTDB-Tk (21) v.0.3.3 implementing the classify_wf command (https://github.com/Ecogenomics/GTDBTk). Briefly, marker genes were identified in each genome, aligned, concatenated and classified with pplacer to identify the maximum-likelihood placement of each genome's concatenated protein alignment in the GTDB-Tk reference tree. GTDB-Tk classifies each genome based on its placement in the reference tree, its relative evolutionary distance and FastANI distance.

Functional annotations

For all MAGs, open reading frames (ORF) were called and annotated using Prokka v.1.13 (22). Additional annotation was performed using the blastp 'verysensitive' setting in Diamond v0.9.24.125 (23) (https://github.com/bbuchfink/diamond.git) against UniRef100 (accessed September 2019) (24), clusters of orthologous groups (COG) (25), Pfam 31 (26) and TIGRfam 15.0 (27). ORFs were assigned KO ID by hmmsearch against KOfam with predefined thresholds (28). Genes of interest were further verified using NCBI's conserved domain search to identify conserved motif(s) present within the gene (29). The genetic potential of each lineages for different metabolism was based on the presence and absence of all key enzymes and a pathway coverage of at least 60% in the KEGG module.

Calculation of MAG relative abundances

To calculate the relative abundance, reads from each metagenomic datasets were mapped to the dereplicated MAGs using CoverM v0.3.1 with the 'contig' command and a cutoff of 95% minimum identity and minimum aligned read length of 75% of each read. Coverage of each contig was calculated with the CoverM 'trimmed_mean' option and the coverage for each MAG was calculated as the average of all contig coverages, weighted by their length. The relative abundance of each MAGs in each metagenomic dataset was calculated as its coverage divided by the total coverage of all genomes in the dereplicate MAGs set.

Determination of MAG diversity and abundance.

To estimate strain level diversity, SingleM v0.12.1 with the 'pipe' command was applied to each raw metagenomic dataset to find abundances of discrete sequence-based operational taxonomic units (OTUs) based on alignment of translated reads to profile HMMs representing conserved ribosomal proteins (https://github.com/wwood/singlem). As 10% of sequences are estimated to be sequencing error, the number of individual strains was calculated as total number of different OTU sequences detected minus 10% of the total sequence count (30). To determine fraction of the microbial community represented by the dereplicated MAG set, read-derived OTU sequences were compared to the ribosomal protein marker genes from the recovered MAGs using the singlem appraise --imperfect --sequence_identity 0.89 command. Read-based OTU sequences with ≥89% global sequence identity were determined to be from the same genus (30).

MAGs that were selectively enriched during the SEP, non-SEP, or Spring-2015 ECF

To determine which lineages were differentially abundant between the SEP, non-SEP, and Spring-2015 ECF periods, the relative abundance profiles of all MAGs were compared across all time points. All statistical analyses were performed using sciPy (31) in the python programming environment (32). The mean and statistical significance of the difference was analyzed using the Mann-Whitney U test. All p-values that were affected by multiple testing were corrected for false discovery using the Benjamini-Hochberg procedure.

Identification of nifH genes from unbinned metagenomic contigs

Metagenomic contigs were annotated with Prodigal using the -p meta option. The ORFs were searched with hmmsearch using the K02588 hmm profile and filtered with the predefined threshold from KOfam (28). The gene sequences were dereplicated with a cutoff of 95% sequence identity using CD-HIT (33). The genes were compared against proteins from GTDB r89 database (34) to infer their taxonomic affiliations.

Read mapping and coverage profile of the nifH gene sequences

Coverage profile of the *nifH* gene sequences from the MAGs and metagenome generated using CoverM v0.3.1 with the 'contig' command and a cutoff of 95% minimum identity and minimum aligned read length of 75% of each read. Coverage of each gene sequence was calculated with the CoverM 'trimmed_mean' option.

Weighted Gene Co-Expression Network Analysis

To identify assemblages of families highly correlating with carbon flux, Weighted Gene

Co-Expression Network Analysis (WGCNA) was performed using the R package WGCNA (35). WGCNA is used in systems biology to identify clusters (modules) of temporally correlated gene expression or taxon abundance profiles, which can be related to external sample traits, in this case carbon flux values. A co-expressed (or co-occurring) network of taxon abundances was represented mathematically using an adjacency matrix that expresses pairwise similarities. Specific modules were then identified by hierarchical clustering. Module associations were explored, by examining first principal component of the module (known as the module eigengene). In our analyses, module and eigengene structure of the WGCNA fitted networks interrelate taxa whose abundances were highly correlated with the magnitude of carbon flux. In combination with module membership and global trait significance values, we identified specific taxonomic groups that were either positively or negatively correlated with carbon flux. For our data, SSU rRNA read counts derived from the metatranscriptome time-series were aggregated by family or higher, if the resolution did not allow for identification at the family level. Then, the dataset was rarefied to the sample with the lowest number of total reads and a transformation was performed using the R "DeSeq2" package variance stabilizing transformation function vst() (11, 36). Families with fewer than 200 reads associated with them were removed from the dataset before further analysis. Modules were defined using a soft-threshold Pearson correlation in conjunction with a topological overlap distance metric and Ward's Hierarchical Agglomerative Clustering Method. Softthresholding powers were chosen using the pickSoftThreshold() WGCNA function. Soft thresholding power was set to 10 and the minimum module size was set at 15 taxa. Visualization of the resulting modules and their correlation with carbon flux was performed using Cytoscape (37).

Random Forest Models

To further define rRNA-based taxon associations with SEP events a Random Forest classification and regression tree machine-learning model was built using the R package randomForest() (38). Using a bagging approach, random forest constructs "forests" of multiple decision trees independently from a bootstrap sample of the dataset. To avoid overfitting, each node, or decision point, was split using a best abundance threshold among a subset of predictors randomly selected rather than using the entire representative dataset. The best split was determined based on the Gini criterion (a variation of entropy measures usually used for decision trees). New data were mapped along the training set trees. Each tree yielded a classification and the majority votes among all the trees were counted and a prediction is assigned. An out-of-bag-error-rate (OOB) was estimated to evaluate the accuracy of classifications. Because our dataset has relatively few samples, partitioning a training set and testing was not possible, and so our trained model was not directly validated on independent data. Instead, we used the mean of squared residuals and percent variance explained as performance metrics. Permutation and cross-validation tests were also performed to assess the accuracy of our model. Due to large differences in their relative abundances, we performed Random Forest models on Eukaryotes and Prokaryotes separately. To decrease dataset noise, rare features were removed (taxonomic groups with less than 200 reads assigned to them) and a z-scale transformation was done before running the model. Our final Prokaryote model was built with 10,001 trees and 32 variables were tried at each split. The OOB estimate was 4.76%, and our permutation test (n = 100) showed the model was significant ($P < 0.05$). The cross-validation test results showed both high accuracy (95.2%) and Kappa (81.4%) values. The eukaryotic Random Forest model had a higher OOB estimate at 14.29%, and lower cross-validation accuracy (85.7%) and Kappa values (15.8%), the permutation test (n = 100)

yielded significant results ($P < 0.05$). Error tables with mean decrease in accuracy of the model for each taxa for both models are provided in Datasets S8 and S9.

Correlation network analyses of sinking particle-associated bacteria and eukaryotes

Correlation networks were inferred between the 2259 families of bacteria, archaea and eukaryotes using FastSpar(v0.0.10) (39), an efficient C++ implementation of the SparCC, which is based on the dbOTU3 algorithm (39-41). SparCC uses a log ratio transformation and iteratively calculates correlations between taxa assuming a sparse network. Correlations between Bacteria, Archaea and Eukaryotes were based on a distribution-based clustering method dbOTU3, 50 iterations and a cut-off threshold of 0.40 ($p < 0.05$) (41). A permutation-based approach was used to assess correlation significance. From the original count data, 1000 bootstrap counts were generated using the command fastspar_bootstrap. From the randomly permuted data, p-values were calculated using fastspar_pvalues. Only interdomain correlation values calculated were retained while correlations between families within prokaryota or eukaryote were eliminated from the analysis. Using the same count table, a Kendall correlation test between taxa and carbon flux values was performed using the cor.test function in the R package (v3.6.0) "Stats"(42). Only correlations with p-values < 0.05 were considered significant. Using the above matrices, insignificant values were eliminated from the data frames. The network was visualized using the Cytoscape (v3.7.2) software to identify the most significant positive correlations between groups and carbon flux (37).

Additionally, Kendall Tau correlation tests were performed on SSU rRNA abundance data, to identify taxa that were specifically enriched during the Spring-2015 ECP (***SI Appendix*** Figure S9; Dataset S9). Only correlation values that were found to be significant ($P = 0.05$) were retained.

Taxa with less than 5,000 SSU rRNA reads across all samples, or those that had a negative Kendall correlation value, were eliminated for visualization purposes (***SI Appendix*** Figure 9).

Data visualization

Figures were generated using pheatmap (43) and ggplot2 (44) package in R (11) and further refined using Adobe Illustrator.

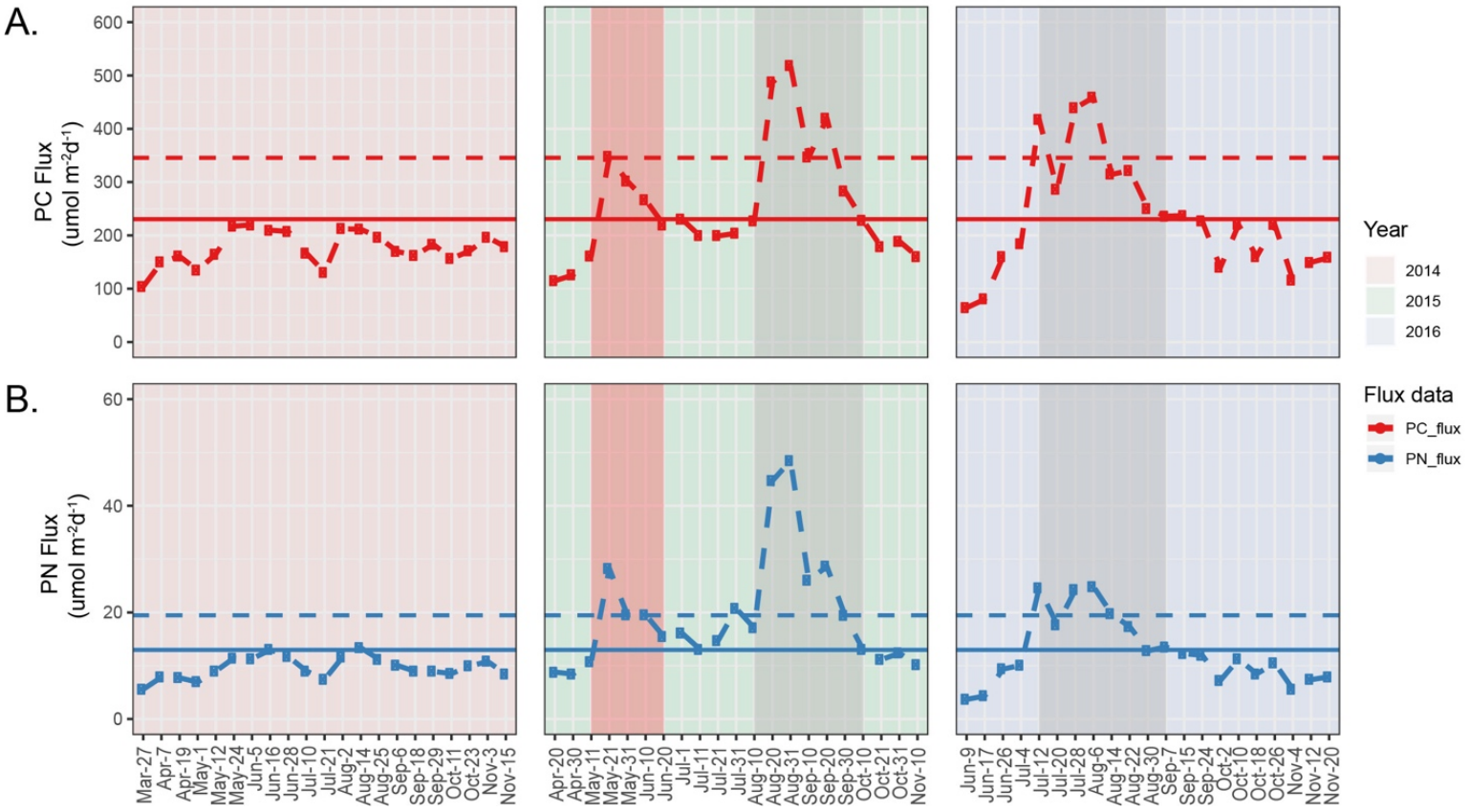

Figure S1.
Total particulate carbon (PC) and nitrogen (PN) fluxes at 4000 m during 2014-2016.
Solid and dashed horizontal lines represent the mean and 150% of the mean annual PC and PN flux values, respectively. Points above the dotted lines represent large flux events. The grey shaded area represents the Summer Export Pulse (SEP), and red shaded area the Spring-2015 ECF event. The values for mean annual PC-flux and PN-flux were 230.4 and 12.96 umol/m^2/d, respectively. Values for PC and PN represent duplicate values in 2014 and 2015 (due to sample availability), and triplicate values for 2016. The average percent difference for all duplicate measurements in 2014 was 4% for PC, and 5%, for PN. The average percent difference for all duplicate measurements in 2015 was 2.9% for PC, and 11.9%, for PN. In 2016, the coefficients of variation from all triplicate measurements for PC ranged from 0 to 0.06, and for PN from 0.01 to 0.17.

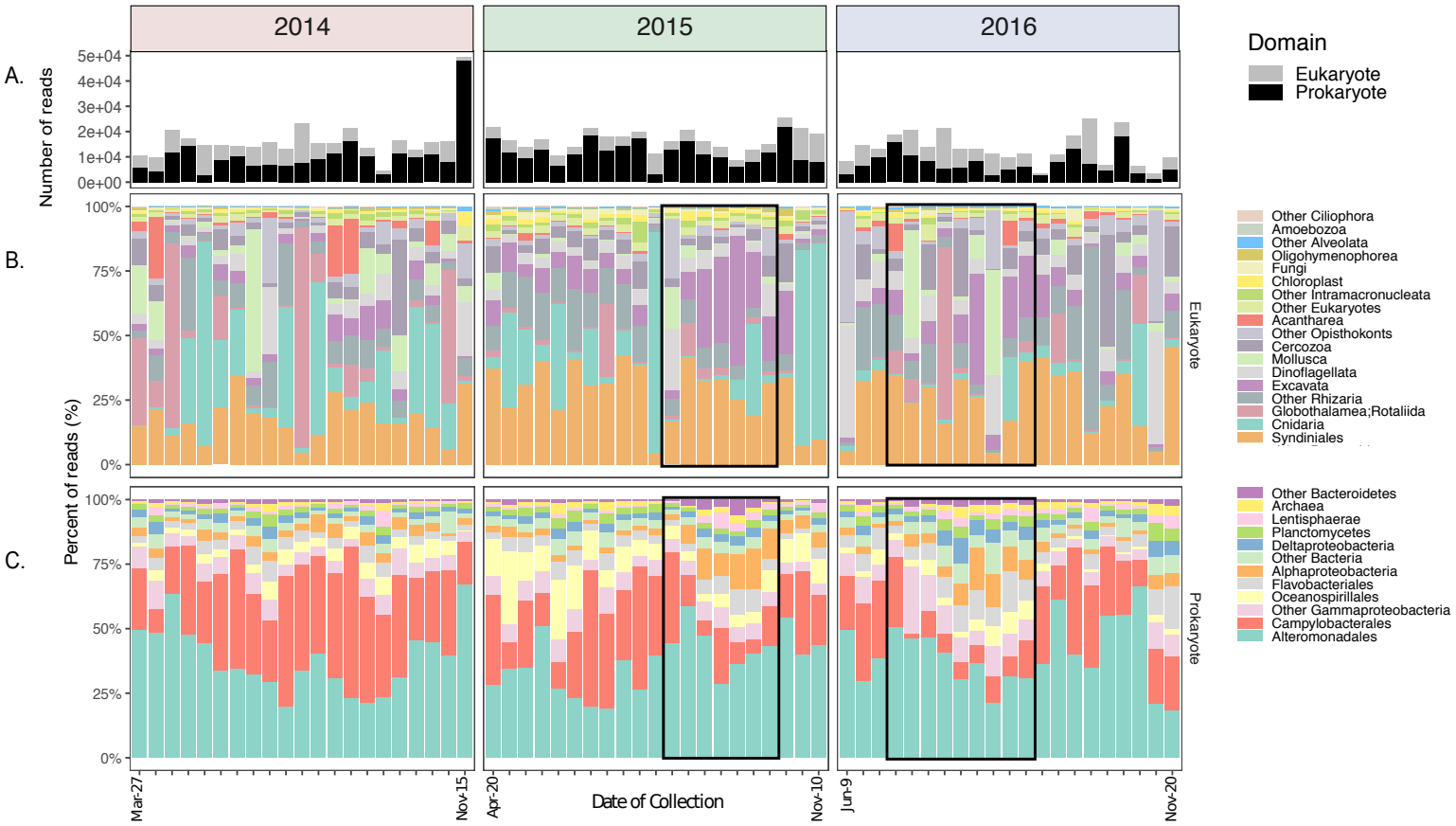

Figure S2. Community profiles based on SSU rRNA gene abundances from metagenomes (DNA) of sinking particles collected at 4,000 m from 2014 -2016.

A. Total small subunit rRNA gene reads recovered from bacteria plus archaea, and eukaryotes. B. Recovery of eukaryotic SSU rRNAs in the sediment trap time-series. C. Recovery of specific bacterial and archaeal SSU rRNAs in the sediment trap time-series. Samples collected during the summer export pulse (SEP) are outlined in black.

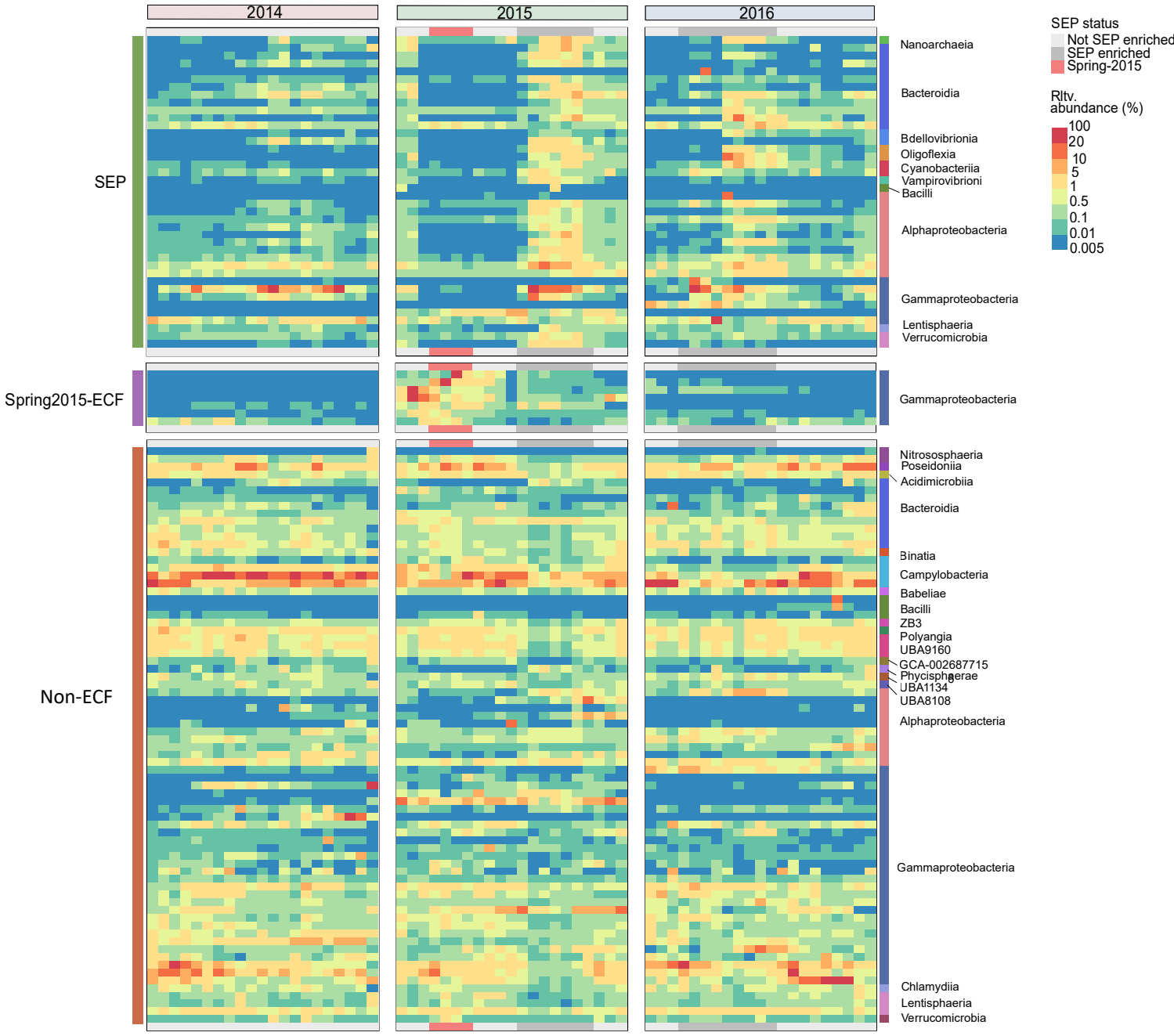

Figure S3. Relative abundance of specific MAGs assembled from metagenomes of sinking particles collected at 4,000 m from 2014 - 2016.

The heat map displays the relative abundance of 121 MAGs determined by read mapping from 63 individual sediment trap sample time points (columns), during 2014 (orange), 2015 (green), and 2016 (blue). Summer export pulse (SEP) time periods are shaded in grey, and Spring-2015 ECF time periods shaded in orange. MAGs significantly enriched ($P < 0.05$) during SEP period are displayed in the top panel, those in the Spring-2015 ECF in the middle panel, and all others (Non-ECF) are displayed in the bottom panel.

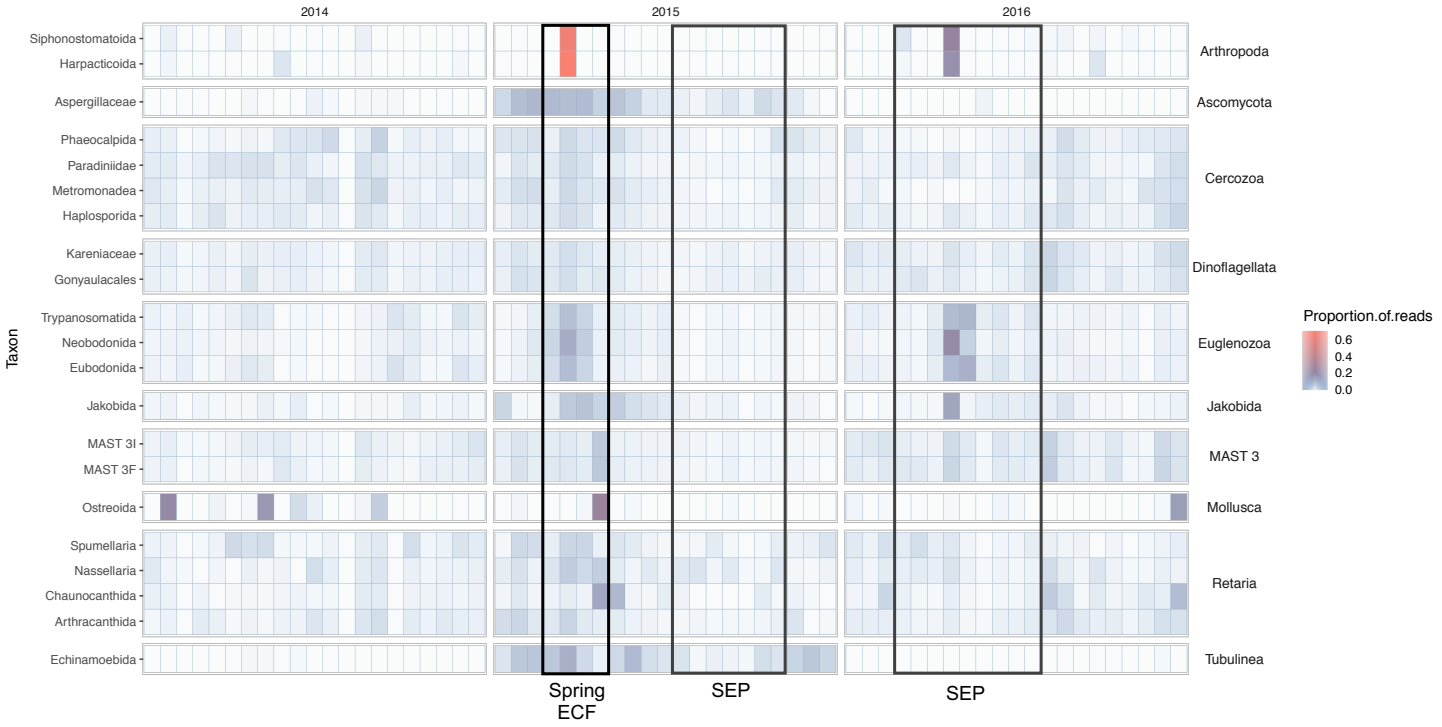

Figure S4. Eukaryotic taxa enriched in rRNA abundance during the Spring-2015 ECF period.

Heat map showing eukaryotic taxa that were significantly enriched in SSU rRNA abundances ($P < 0.05$) based on Kendall's method during the Spring-2015 ECF period. Values were calculated proportionately across rows. Samples collected during the ECF periods are outlined in dark black (Spring-2015) or light black (SEP periods in 2015 and 2016). Only taxa with >5000 reads across the time series are shown.

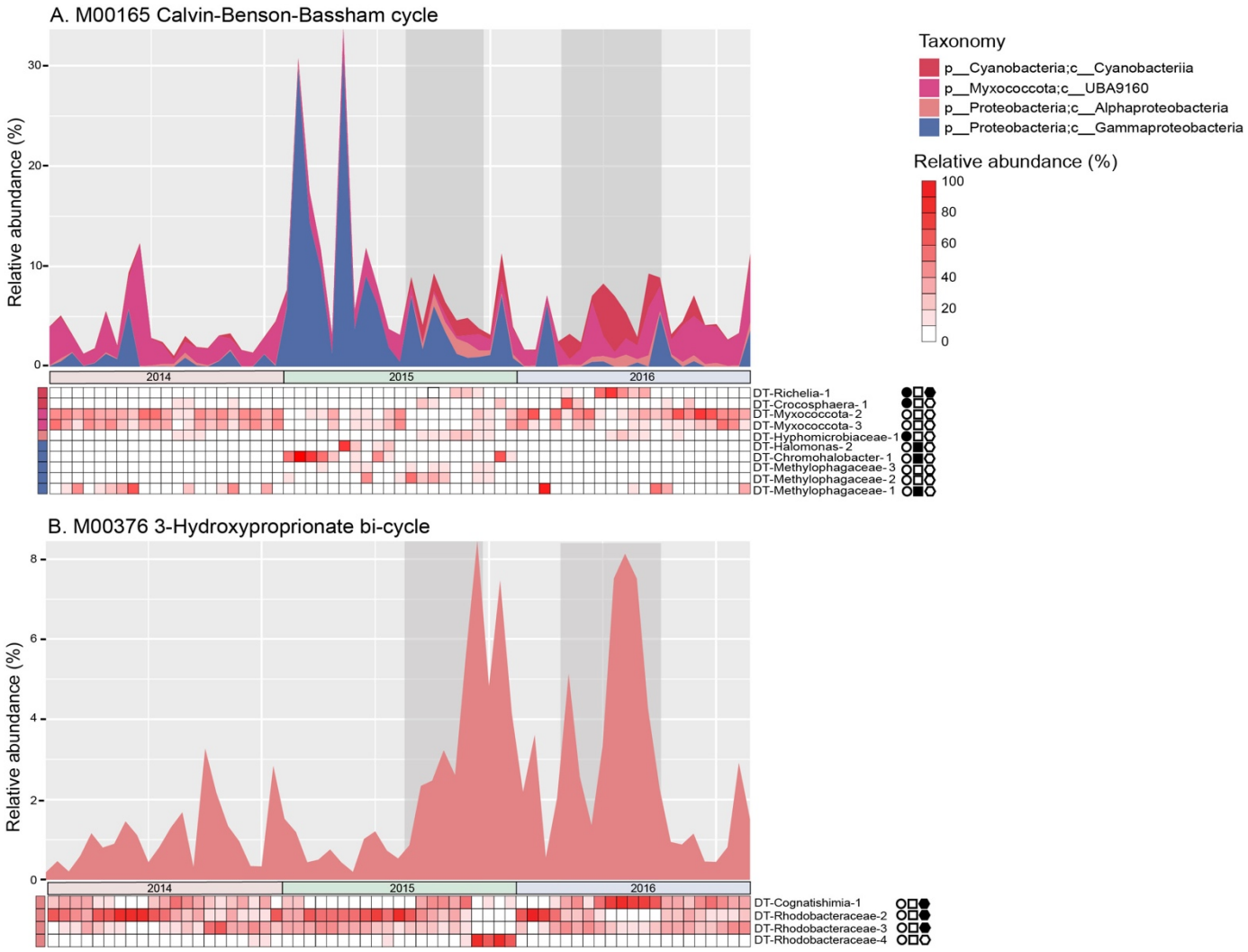

Figure S5. Changes in the abundance of MAGs that encoded carbon fixation genes on sinking particles collected at 4,000 m from 2014 - 2016.

Key metabolic pathways encoded by different MAGs driving carbon fixation. All MAGs found to encode these pathways are shown in the plots. A**.** Calvin-Benson-Bassham cycle and B. 3-Hydroxyproprionate bi-cycle. The area chart represents cumulative abundance of the MAGs contributing to the specific metabolic process. The heatmap represents the relative contribution of the individual MAG to the total abundance of the specific metabolic process. The symbols in solid black to the right of the heatmap indicate those MAGs that positively correlated with either the SEP time period (circles), the Spring-2015 ECF (squares), or carbon flux (polygon). Empty symbols indicate no correlation found. The time periods of the summer export pulse (SEP) are shaded in darker grey.

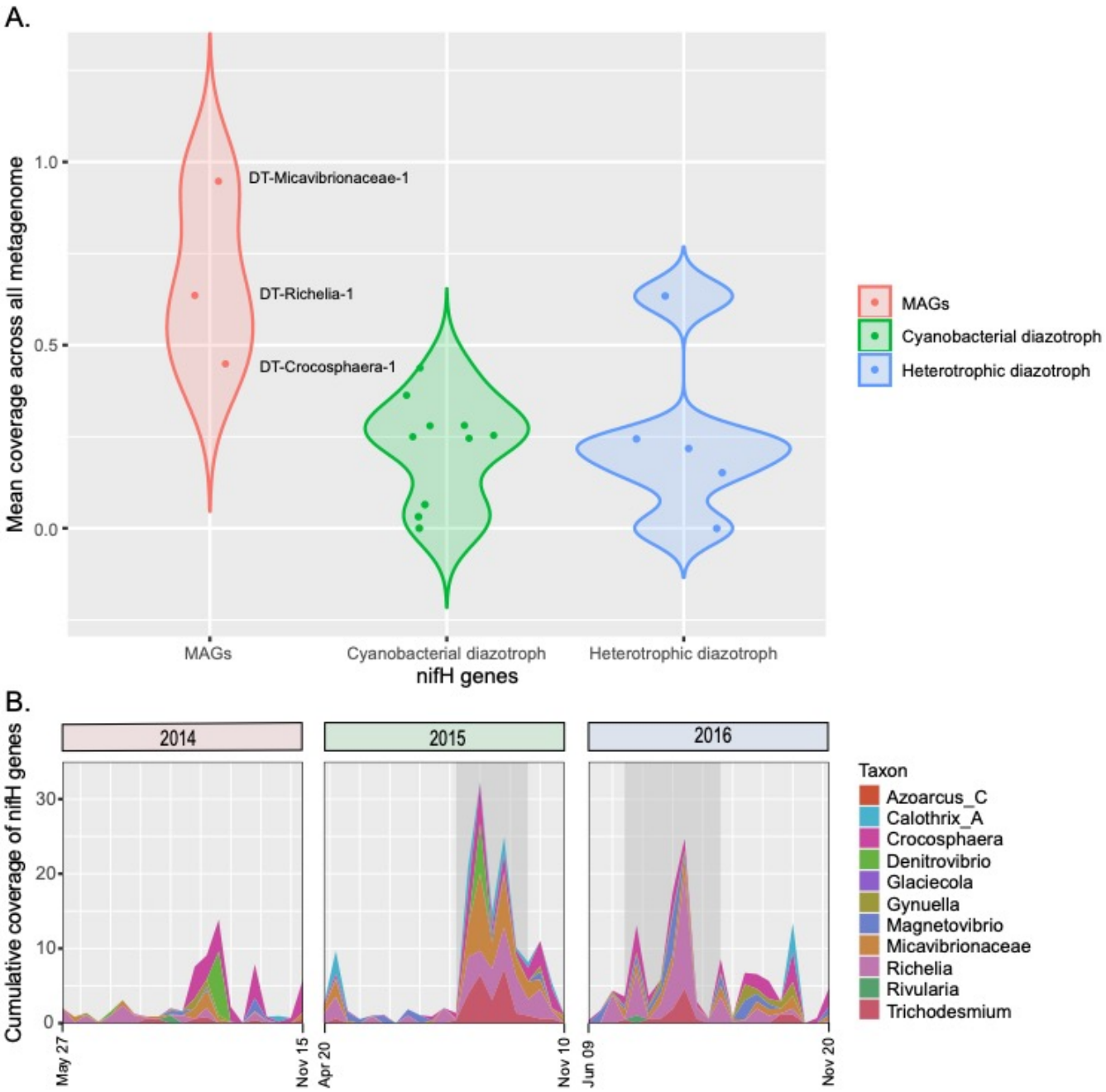

Figure S6. Relative abundance of *nifH* genes recovered from metagenomes of sinking particles collected at 4,000 m from 2014 - 2016.

A. Violin plots summarizing the average mean coverage of *nifH* genes recovered from MAGs and orphan contigs across all metagenomic samples. The *nifH* genes were separated into three groups (MAGs, diazotrophic, and heterotrophic orphan genes) (See Supplementary Table 5 for top blast hits and coverage values of the orphan genes). Coverage values were calculated after removing the 5% of nucleotide positions with the highest and lowest coverage. B. Area chart representing the cumulative coverage values of the *nifH* genes by taxa across the three year metagenomic time series. The time periods of the summer export pulse (SEP) are shaded in darker grey.

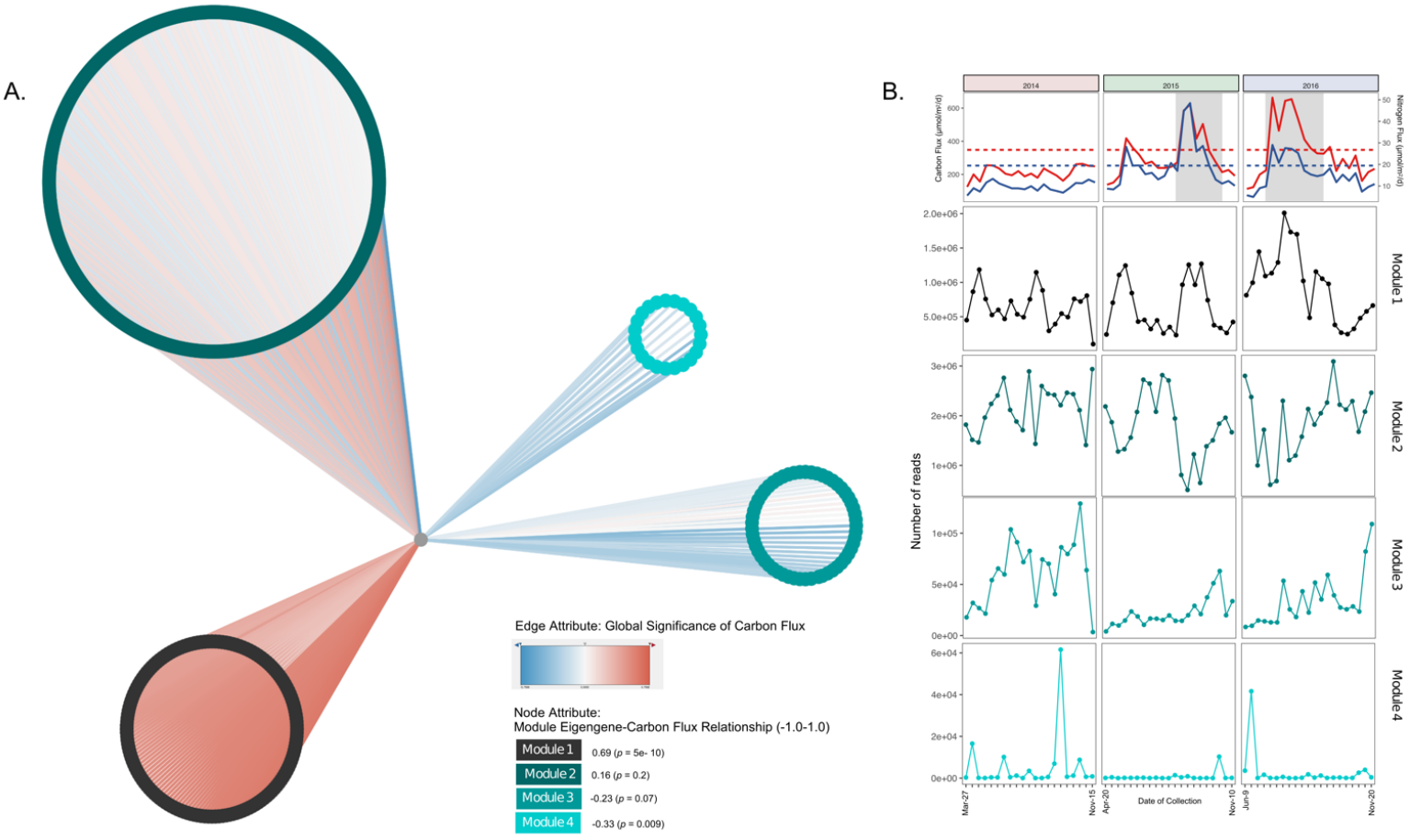

Figure S7. Correlations of rRNA-based taxon abundances with carbon-flux via WGCNA analyses.

A. Eigengenes representing the relative abundance of four modules identified in a weighted network analysis are shown with module eigengene to carbon flux relationship denoted by node color for each of the four modules. The edge color represents individual taxon global significance of carbon flux (central grey node) from most negative to most positive (blue to red). The size of the nodes represents the cumulative relative abundance in the module.

B. The cumulative relative abundances of taxa in each of the four WGCNA modules is plotted below the corresponding carbon flux values (red line) and nitrogen flux values (blue line). Module 1 positively correlates with elevated carbon flux. See Dataset S4 for the taxon composition of each WGCNA module and Methods for more information.

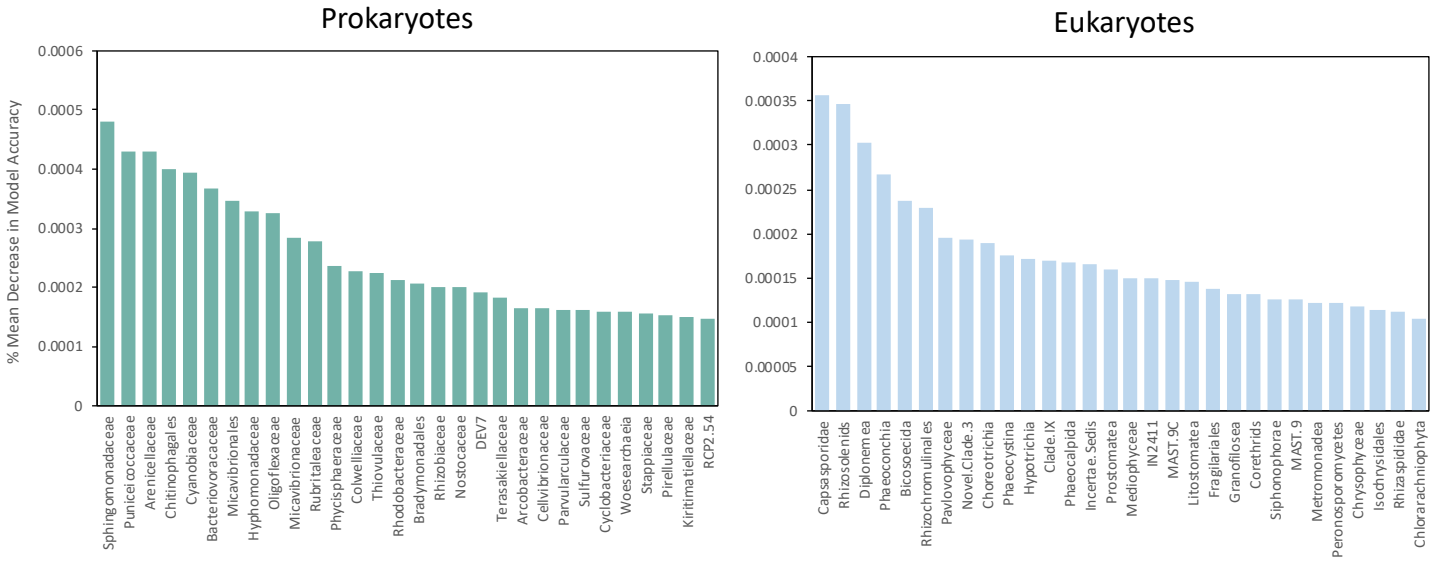

Figure S8. Random Forest Mean decrease in accuracy ranking for Prokaryote (Bacteria and Archaea) or Eukaryote models.

Left panel: Top 30 bacterial and archaeal taxa ranked by mean decrease in accuracy for a Random Forest classification model run using bacterial and archaeal SSU rRNAs. For the models, a total of 10,001 trees were built. Left Panel:Top prokaryotic taxa ranked by mean decrease in accuracy for a Random Forest classification model run using Prokaryote SSU rRNAs. Right panel: Top 30 eukaryotic taxa ranked by mean decrease in accuracy for a Random Forest classification model run using Eukaryote SSU rRNAs. Specific details on model testing are further provided in the Methods.

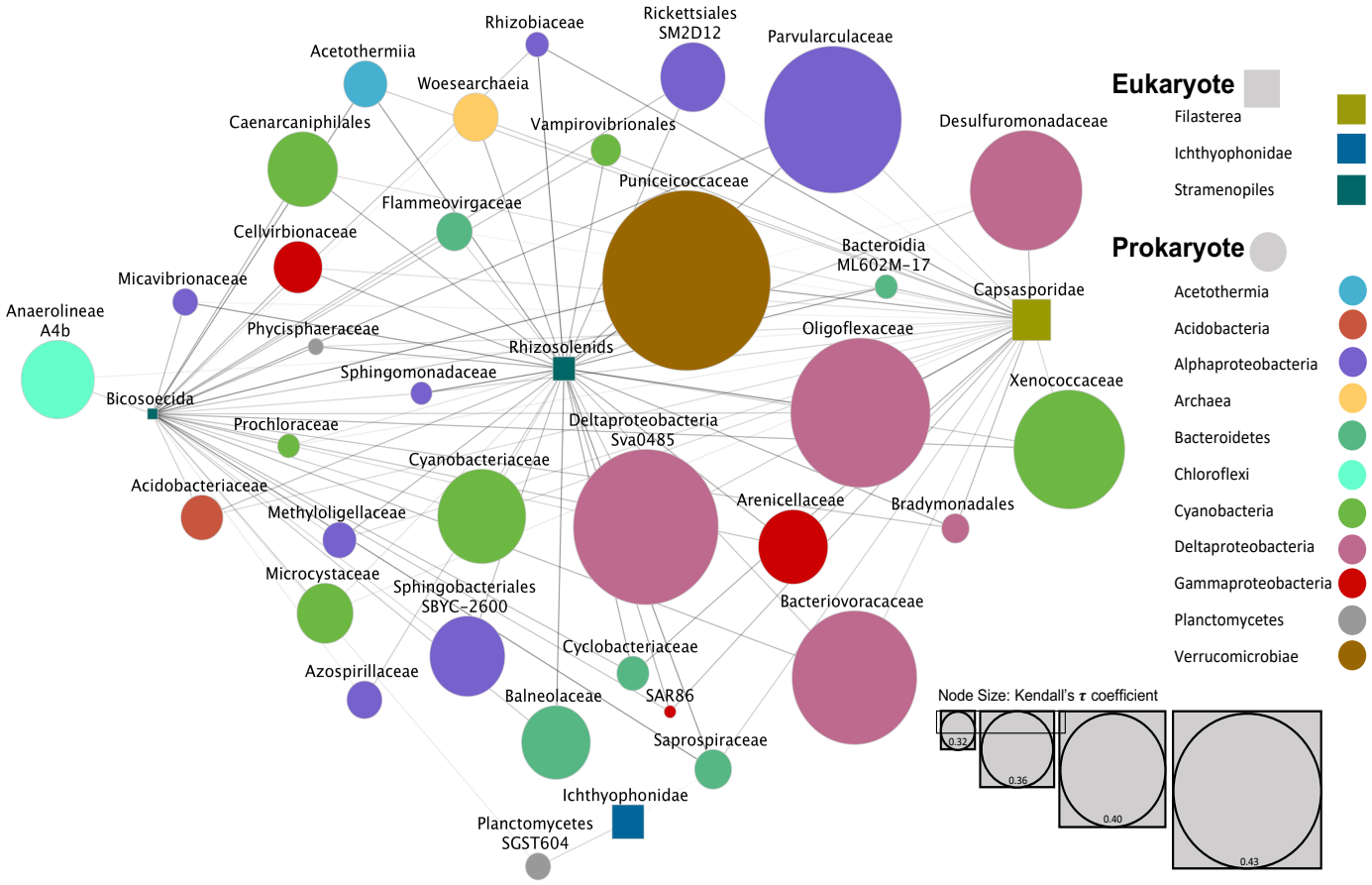

Figure S9. Prokaryote-Eukaryote network associations determined from SSU rRNA abundance correlations to carbon flux.

Correlations between Bacteria, Archaea and Eukaryotes were based on a distribution-based clustering method dbOTU3 (39), 50 iterations and a cut-off threshold of 0.40 ($p < 0.05$). Kendall's Tau coefficient calculated between taxonomic groups and carbon flux values are denoted by the node size, with a cutoff threshold of 0.30 assigned ($p < 0.05$). The larger the node size, the greater the positive correlation of the taxon with carbon flux. Nodes are colored by higher taxonomic level and Prokaryotes and Eukaryotes assigned by node shape.

Family	Genus	Percent of total reads (MT)	Percent of Campylobacterales reads (MT)	Percent of total reads (MG)	Percent of Campylobacterales reads (MG)
Arcobacteraceae	*Arcobacter*	34.31%	99.37%	16.36%	99.37%
Sulfurovaceae	*Sulfurovum*	0.15%	0.45%	<0.01%	0.45%
Sulfurovaceae	*Nitratifractor*	<0.01%	<0.01%	<0.01%	<0.01%
Thiovulaceae	*Sulfurimonas*	0.03%	0.09%	<0.01%	0.09%
Thiovulaceae	*Sulfuricurvum*	0.01%	0.04%	<0.01%	0.04%
Thiovulaceae	*Thiovulum*	<0.01%	0.01%	<0.01%	0.01%
Sulfurospirillaceae	*Sulfurospirillum*	0.01%	0.04%	0.01%	0.04%
Helicobacteraceae	*uncultured*	0.01%	0.06%	0.01%	0.06%
Campylobacteraceae	*Campylobacter*	<0.01%	<0.01%	0.08%	<0.01%
Nitratiruptoraceae	*Hydrogenimonas*	<0.01%	<0.01%	0.01%	<0.01%
Other		<0.01%	<0.01%	<0.01%	<0.01%
Total		34.53%	100%	16.47%	100%

Table S1. Top 10 most abundant Campylobacterales genera represented in the SSU rRNA identified from the sediment trap metatranscriptomes (MT), organized by percent of total reads in each family. Respective metagenomic read percentages (MG) are also shown.

Order	Family	Genus	Percent of total reads (MT)	Percent of Gammaproteobacteria reads (MT)	Percent of total reads (MG)	Percent of Gammaproteobacteria reads (MG)
Alteromonadales	Colwelliaceae	*Colwellia*	17.94%	45.81%	10.87%	32.46%
Alteromonadales	Moritellaceae	*Moritella*	6.57%	16.76%	7.92%	23.65%
Alteromonadales	Shewanellaceae	*Psychrobium*	1.73%	4.43%	1.51%	4.50%
Alteromonadales	Shewanellaceae	*Shewanella*	1.07%	2.73%	2.71%	8.10%
Alteromonadales	Pseudoalteromonadaceae	*Pseudoalteromonas*	1.02%	2.60%	0.75%	2.23%
Alteromonadales	Colwelliaceae	*Thalassotalea*	0.57%	1.44%	0.35%	1.05%
Alteromonadales	Pseudoalteromonadaceae	-	0.48%	1.24%	0.11%	0.33%
Alteromonadales	Colwelliaceae	-	0.46%	1.18%	0.20%	0.59%
Alteromonadales	Psychromonadaceae	*Psychromonas*	0.33%	0.84%	0.29%	0.86%
Alteromonadales	Alteromonadaceae	*Alteromonas*	0.16%	0.42%	0.12%	0.36%
Alteromonadales	Moritellaceae	*Paramoritella*	0.16%	0.40%	0.22%	0.65%
Alteromonadales	Shewanellaceae	*Ferrimonas*	0.09%	0.24%	0.06%	0.17%
Alteromonadales	Psychromonadaceae	*Agarivorans*	0.07%	0.19%	0.01%	0.03%
Alteromonadales	Idiomarinaceae	*Idiomarina*	0.06%	0.16%	0.39%	1.16%
Oceanospirillales	Nitrincolaceae	*Profundimonas*	3.08%	7.86%	1.96%	5.87%
Oceanospirillales	Oleiphilaceae	*Oleiphilus*	0.49%	1.24%	0.09%	0.27%
Oceanospirillales	Saccharospirillaceae	*Oleispira*	0.47%	1.20%	0.15%	0.44%
Oceanospirillales	Saccharospirillaceae	-	0.35%	0.90%	0.13%	0.39%
Oceanospirillales	Kangiellaceae	*Aliikangiella*	0.33%	0.85%	0.11%	0.34%
Oceanospirillales	Saccharospirillaceae	-	0.26%	0.67%	0.04%	0.13%
Oceanospirillales	Nitrincolaceae	*uncultured*	0.15%	0.38%	0.08%	0.23%
Oceanospirillales	Marinomonadaceae	*Marinomonas*	0.07%	0.19%	0.06%	0.19%
Vibrionales	Vibrionaceae	*Vibrio*	0.36%	0.91%	0.48%	1.45%
Vibrionales	Vibrionaceae	*Photobacterium*	0.28%	0.70%	0.17%	0.51%
Vibrionales	Vibrionaceae	*Enterovibrio*	0.07%	0.19%	0.01%	0.02%
Cellvibrionales	Spongiibacteraceae	*BD1-7 clade*	0.22%	0.56%	0.23%	0.69%
Cellvibrionales	Halieaceae	*Halioglobus*	0.21%	0.54%	0.23%	0.67%
Cellvibrionales	Cellvibrionaceae	*uncultured*	0.12%	0.31%	0.05%	0.16%
Cellvibrionales	Halieaceae	*OM60(NOR5) clade*	0.07%	0.18%	0.15%	0.44%
Cellvibrionales	Spongiibacteraceae	*Sinobacterium*	0.07%	0.17%	0.03%	0.09%
Nitrosococcales	Methylophagaceae	*Methylophaga*	0.10%	0.26%	0.08%	0.23%
Steroidobacterales	Woeseiaceae	*Woeseia*	0.09%	0.22%	0.15%	0.45%
Other			1.67%	4.23%	3.79%	11.29%
Total			39.17%	100%	33.50%	100%

Table S2. Top 32 most abundant Gammaproteobacteria genera represented in the SSU rRNA identified from the sediment trap metatranscriptomes (MT) and organized by percent of total reads in each order. Respective metagenomic read percentages (MG) are also shown. -, undesignated

Order	Family	Genus	Percent of reads (MT)	Percent of Alphaproteobacteria reads (MT)	Percent of reads (MG)	Percent of Alphaproteobacteria reads (MG)
Rhodobacterales	Rhodobacteraceae	*Tropicibacter*	0.08%	4.35%	0.15%	5.61%
Rhodobacterales	Rhodobacteraceae	*Sedimentitalea*	0.06%	3.55%	0.13%	4.82%
Rhodobacterales	Rhodobacteraceae	*Thalassobius*	0.05%	2.66%	0.12%	4.61%
Rhodobacterales	Rhodobacteraceae	*Ruegeria*	0.04%	2.20%	0.13%	5.05%
Rhodobacterales	Rhodobacteraceae	*Roseovarius*	0.03%	1.74%	0.03%	1.15%
Rhodobacterales	Rhodobacteraceae	*Loktanella*	0.03%	1.71%	0.04%	1.60%
Rhodobacterales	Rhodobacteraceae	*Leisingera*	0.03%	1.50%	0.04%	1.56%
Rhodobacterales	Rhodobacteraceae	*Aestuariivita*	0.02%	1.34%	0.03%	1.21%
Rhodobacterales	Rhodobacteraceae	*Phaeobacter*	0.02%	1.21%	0.02%	0.85%
Rhodobacterales	Rhodobacteraceae	*Shimia*	0.02%	0.90%	0.03%	1.30%
Rhodobacterales	Rhodobacteraceae	*Pseudophaeobacter*	0.02%	0.87%	0.03%	1.08%
Rhodobacterales	Rhodobacteraceae	*Dinoroseobacter*	0.02%	0.83%	0.02%	0.63%
Rhodobacterales	Rhodobacteraceae	*Sulfitobacter*	0.01%	0.80%	0.03%	0.97%
Rhodobacterales	Rhodobacteraceae	*Actibacterium*	0.01%	0.80%	0.01%	0.53%
Rhodobacterales	Rhodobacteraceae	*Thalassococcus*	0.01%	0.72%	0.02%	0.58%
Rhodobacterales	Rhodobacteraceae	*HIMB11*	0.01%	0.68%	0.01%	0.52%
Rhodobacterales	Rhodobacteraceae	*Roseobacter*	0.01%	0.59%	0.01%	0.33%
Rhodobacterales	Rhodobacteraceae	*Tropicimonas*	0.01%	0.56%	0.01%	0.35%
Caulobacterales	Hyphomonadaceae	-	0.22%	12.29%	0.14%	5.35%
Caulobacterales	Parvularculaceae	*Parvularcula*	0.05%	2.65%	0.01%	0.53%
Caulobacterales	Hyphomonadaceae	*Maricaulis*	0.03%	1.89%	0.02%	0.65%
Caulobacterales	Hyphomonadaceae	*Oceanicaulis*	0.02%	0.92%	0.01%	0.33%
Rhizobiales	Rhizobiaceae	*Pseudahrensia*	0.05%	2.75%	0.04%	1.42%
Rhizobiales	Hyphomicrobiaceae	*Filomicrobium*	0.02%	1.25%	0.03%	0.96%
Rhizobiales	Stappiaceae	*Labrenzia*	0.02%	1.22%	0.05%	1.72%
Rhizobiales	Rhizobiaceae	*Lentilitoribacter*	0.01%	0.74%	0.02%	0.92%
Sphingomonadales	Sphingomonadaceae	*Erythrobacter*	0.09%	5.20%	0.10%	3.71%
Sphingomonadales	Sphingomonadaceae	*Altererythrobacter*	0.02%	1.14%	0.04%	1.46%
Parvibaculales	PS1 clade	*Pseudovibri*	0.03%	1.57%	0.08%	2.85%
Rickettsiales	Rickettsiaceae	*Candidatus Megaira*	0.03%	1.44%	0.01%	0.36%
Other			0.95%	39.92%	1.25%	46.98%
Total			2.05%	100%	2.66%	100%

-,

Table S3. Top 30 most abundant Alphaproteobacteria genera represented in the SSU rRNA identified from the sediment trap metatranscriptomes (MT) and organized by percent of total reads in each order. Respective metagenomic read percentages (MG) are also shown. -, undesignated

Order	Family	Percent of reads (MT)	Percent of Deltaproteobacteria reads (MT)	Percent of reads (MG)	Percent of Deltaproteobacteria reads (MG)
NB1-j clade	-	0.35%	28.63%	0.70%	43.64%
Myxococcales	-	0.18%	14.31%	0.11%	6.77%
Myxococcales	Sandaracinaceae	0.10%	7.90%	0.05%	2.88%
Myxococcales	Nannocystaceae	0.06%	5.06%	0.07%	3.99%
Myxococcales	Haliangiaceae	0.02%	1.57%	0.01%	0.79%
Myxococcales	-	0.02%	1.59%	0.02%	1.44%
Bdellovibrionales	Bacteriovoracaceae	0.16%	13.23%	0.11%	7.01%
Bdellovibrionales	Bdellovibrionaceae	0.12%	9.98%	0.12%	7.47%
Oligoflexales	Oligoflexaceae	0.05%	4.43%	0.15%	9.14%
PB19 clade	-	0.03%	3.20%	0.04%	2.44%
Desulfarculales	Desulfarculaceae	0.01%	1.06%	0.03%	2.10%
Desulfobacterales	Desulfobacteraceae	0.01%	0.85%	0.02%	1.01%
RCP2-54 clade	-	0.01%	0.73%	0.02%	1.14%
SAR324 clade	-	0.01%	0.50%	0.02%	1.25%
Bradymonadales	-	0.01%	0.49%	NA	NA
Other		0.09%	6.47%	0.15%	8.93%
Total		1.23%	100%	1.63%	100%

Table S4. Top 20 most abundant Deltaproteobacteria families represented in the SSU rRNA identified from the sediment trap metatranscriptomes (MT) and organized by percent of total reads in each order. Respective metagenomic read percentages (MG) are also shown. -, undesignated

Class	Order	Family	Percent of reads (MT)	Percent of Bacteroidetes reads (MT)	Percent of reads (MG)	Percent of Bacteroidetes reads (MG)
Bacteroidia	Flavobacteriales	Flavobacteriaceae	1.13%	31.37%	1.62%	45.55%
Bacteroidia	Flavobacteriales	Crocinitomicaceae	0.84%	23.14%	0.59%	16.62%
Bacteroidia	Flavobacteriales	Cryomorphaceae	0.59%	16.32%	0.31%	8.83%
Bacteroidia	Flavobacteriales	NS9 marine group	0.08%	2.19%	0.26%	7.40%
Bacteroidia	Flavobacteriales	NS7 marine group	0.03%	0.77%	0.02%	0.58%
Bacteroidia	Flavobacteriales	Weeksellaceae	<0.01%	0.09%	0.01%	0.26%
Bacteroidia	Flavobacteriales	Schleiferiaceae	<0.01%	0.05%	<0.01%	0.13%
Bacteroidia	Cytophagales	Cyclobacteriaceae	0.40%	11.09%	0.30%	8.49%
Bacteroidia	Cytophagales	Flammeovirgaceae	0.02%	0.59%	0.01%	0.33%
Bacteroidia	Cytophagales	Amoebophilaceae	0.01%	0.18%	0.02%	0.43%
Bacteroidia	Cytophagales	Cytophagaceae	<0.01%	0.12%	<0.01%	0.09%
Bacteroidia	Cytophagales	Microscillaceae	<0.01%	0.09%	<0.01%	0.10%
Bacteroidia	Cytophagales	-	<0.01%	0.07%	<0.01%	0.04%
Bacteroidia	Cytophagales	Bernardetiaceae	<0.01%	0.03%	<0.01%	0.05%
Bacteroidia	Cytophagales	Hymenobacteraceae	<0.01%	0.03%	<0.01%	0.03%
Bacteroidia	Chitinophagales	Saprospiraceae	0.37%	10.26%	0.27%	7.58%
Bacteroidia	Chitinophagales	-	0.03%	0.76%	0.01%	0.32%
Bacteroidia	Chitinophagales	-	0.02%	0.52%	0.01%	0.29%
Bacteroidia	Chitinophagales	-	0.02%	0.43%	0.01%	0.26%
Bacteroidia	Chitinophagales	-	<0.01%	0.14%	<0.01%	0.11%
Bacteroidia	Chitinophagales	Chitinophagaceae	<0.01%	0.03%	<0.01%	0.05%
Bacteroidia	Sphingobacteriales	Lentimicrobiaceae	0.01%	0.27%	0.01%	0.37%
Bacteroidia	Sphingobacteriales	-	0.01%	0.18%	0.01%	0.18%
Bacteroidia	Sphingobacteriales	NS11-12 marine group	<0.01%	0.11%	<0.01%	0.09%
Bacteroidia	Sphingobacteriales	-	<0.01%	0.07%	<0.01%	0.01%
Bacteroidia	Sphingobacteriales	-	<0.01%	0.07%	<0.01%	0.06%
Bacteroidia	Sphingobacteriales	-	<0.01%	0.06%	<0.01%	0.01%
Bacteroidia	Sphingobacteriales	-	<0.01%	0.05%	<0.01%	0.09%
Bacteroidia	Bacteroidales	Prolixibacteraceae	<0.01%	0.07%	<0.01%	0.06%
Bacteroidia	Bacteroidales	-	<0.01%	0.04%	<0.01%	0.03%
Bacteroidia	Bacteroidales	-	<0.01%	0.04%	<0.01%	0.01%
Bacteroidia	Bacteroidales	Bacteroidetes BD2-2	<0.01%	0.04%	<0.01%	0.07%
Bacteroidia	Bacteroidales	Marinifilaceae	<0.01%	0.03%	0.01%	0.24%
Bacteroidia	Bacteroidales	-	<0.01%	0.03%	<0.01%	<0.01%
Bacteroidia	ML602M-17	marine metagenome	<0.01%	0.11%	<0.01%	<0.01%
Rhodothermia	Rhodothermales	Rhodothermaceae	<0.01%	0.10%	0.01%	0.35%
Rhodothermia	Balneolales	Balneolaceae	0.01%	0.18%	0.01%	0.29%
Other			0.01%	0.31%	0.03%	0.58%
Total			3.62%	100%	3.56%	100%

Table S5. Top 31 most abundant Bacteroidetes families represented in the SSU rRNA identified from the sediment trap metatranscriptomes (MT) and organized by percent of total reads in each class. Respective metagenomic read percentages (MG) are also shown. -, undesignated

Supergroup	Phylum	Class	Percent of reads (MT)	Percent of SAR reads (MT)	Percent of reads (MG)	Percent of SAR reads (MG)
Alveolata	Ciliophora	Intramacronucleata	3.61%	42.30%	0.52%	2.39%
Alveolata	Protalveolata	Syndiniales	0.50%	5.89%	7.89%	35.88%
Alveolata	Dinoflagellata	Dinophyceae	0.29%	3.45%	0.71%	3.25%
Alveolata	uncultured	uncultured eukaryote	0.12%	1.35%	0.26%	0.06%
Alveolata	Ciliophora	Postciliodesmatophora	0.05%	0.54%	0.04%	0.20%
Rhizaria	Cercozoa	Ascetosporea	1.53%	17.92%	0.47%	2.13%
Rhizaria	Retaria	Acantharia	0.31%	3.60%	0.95%	4.32%
Rhizaria	Cercozoa	Imbricatea	0.26%	3.02%	0.07%	0.30%
Rhizaria	Cercozoa	Phaeocalpida	0.22%	2.59%	0.54%	2.43%
Rhizaria	Cercozoa	Phaeogromia	0.20%	2.31%	0.54%	2.45%
Rhizaria	Retaria	Foraminifera	0.18%	2.13%	4.83%	21.99%
Rhizaria	Cercozoa	Thecofilosea	0.16%	1.89%	0.16%	0.73%
Rhizaria	Cercozoa	uncultured	0.16%	1.85%	0.26%	0.06%
Rhizaria	Cercozoa	uncultured	0.09%	1.09%	0.18%	0.04%
Rhizaria	Retaria	RAD B	0.09%	1.03%	0.14%	0.65%
Rhizaria	Retaria	RAD C	0.06%	0.66%	0.20%	0.93%
Rhizaria	Retaria	Polycystinea	0.05%	0.61%	3.57%	16.23%
Rhizaria	Cercozoa	Cyranomonas	0.04%	0.52%	0.02%	0.07%
Rhizaria	Cercozoa	Phaeoconchia	0.04%	0.48%	0.05%	0.24%
Rhizaria	Cercozoa	Novel Clade 10	0.04%	0.46%	0.03%	0.14%
Rhizaria	Cercozoa	Cercomonadidae	0.04%	0.44%	0.02%	0.09%
Rhizaria	Cercozoa	Novel Clade Gran-3	0.03%	0.35%	0.04%	0.18%
Rhizaria	Cercozoa	Granofilosea	0.02%	0.26%	0.01%	0.04%
Rhizaria	Cercozoa	uncultured	0.02%	0.22%	0.06%	0.01%
Rhizaria	Cercozoa	Gromia	0.02%	0.20%	<0.01%	0.02%
Rhizaria	Cercozoa	Phaeodina sp. Go1	0.02%	0.20%	0.02%	0.07%
Rhizaria	Cercozoa	Glissomonadida	0.02%	0.19%	0.01%	0.06%
Rhizaria	Cercozoa	Metromonadea	0.01%	0.17%	<0.01%	0.01%
Stramenopiles	Ochrophyta	Diatomea	0.09%	1.10%	0.06%	0.29%
Stramenopiles	MAST-3	MAST-3I	0.03%	0.31%	0.01%	0.03%
Stramenopiles	Labyrinthulomycetes	Thraustochytriaceae	0.02%	0.24%	0.03%	0.12%
Other			0.23%	2.64	0.29%	4.58%
Total			8.54%	100%	21.98%	100%

Table S6. Top 31 most abundant SAR (Stramenopiles, Alveolates, and Rhizaria) classes represented in the SSU rRNA identified from the sediment trap metatranscriptomes (MT) and organized by percent of total reads in each supergroup. Respective metagenomic read percentages (MG) are also shown.

Phylum	Class	Subclass	Percent of reads (MT)	Percent of Metazoa reads (MT)	Percent of reads (MG)	Percent of Metazoa reads (MG)
Mollusca	Gastropoda	Heterobranchia	2.57%	47.39%	1.92%	17.07%
Mollusca	Gastropoda	Caenogastropoda	0.14%	2.62%	0.05%	0.46%
Mollusca	Bivalvia	Pteriomorphia	0.10%	1.90%	0.24%	2.16%
Mollusca	Gastropoda	Neritimorpha	0.02%	0.37%	<0.01%	0.03%
Mollusca	Bivalvia	Heteroconchia	0.01%	0.19%	<0.01%	0.04%
Mollusca	Bivalvia	Palaeoheterodonta	0.01%	0.10%	<0.01%	0.01%
Mollusca	Bivalvia	Protobranchia	<0.01%	0.04%	<0.01%	0.04%
Mollusca	Cephalopoda	Coleoidea	<0.01%	0.04%	0.04%	0.36%
Cnidaria	Hydrozoa	Hydroidolina	1.57%	28.92%	6.72%	59.78%
Cnidaria	Hydrozoa	Trachylinae	0.24%	4.35%	0.15%	1.38%
Cnidaria	Scyphozoa	Semaeostomeae	<0.01%	0.04%	<0.01%	0.01%
Cnidaria	Scyphozoa	Rhizostomeae	<0.01%	0.03%	<0.01%	0.01%
Ctenophora	Tentaculata	Cydippida	0.23%	4.16%	0.04%	0.31%
Ctenophora	Nuda	Beroida	0.03%	0.57%	<0.01%	0.03%
Ctenophora	Tentaculata	Lobata	0.01%	0.17%	0.01%	0.06%
Ctenophora	Tentaculata	Platyctenida	0.01%	0.13%	<0.01%	0.04%
Ctenophora	Tentaculata	Lyrocteis sp. LMC-2016	<0.01%	0.08%	<0.01%	0.02%
Ctenophora	Tentaculata	Cestida	<0.01%	0.07%	<0.01%	0.02%
Ctenophora	Tentaculata	uncultured eukaryote	<0.01%	0.05%	0.01%	0.11%
Chordata	Tunicata	Appendicularia	0.17%	3.10%	0.11%	1.00%
Chordata	Vertebrata	Gnathostomata	0.01%	0.23%	0.28%	2.46%
Chordata	Tunicata	Thaliacea	0.01%	0.12%	0.38%	3.37%
Arthropoda	Crustacea	Maxillopoda	0.14%	2.58%	0.19%	1.70%
Arthropoda	Crustacea	Malacostraca	0.01%	0.15%	0.03%	0.24%
Arthropoda	Hexapoda	Insecta	0.01%	0.10%	0.01%	0.09%
Arthropoda	Crustacea	Ostracoda	0.01%	0.09%	0.04%	0.36%
Annelida	Polychaeta	Palpata	0.07%	1.29%	0.24%	2.16%
Annelida	-	-	0.02%	0.29%	NA	NA
Annelida	Polychaeta	Scolecida	0.02%	0.29%	0.48%	4.27%
Annelida	-	-	<0.01%	0.05%	NA	NA
Other			0.02%	0.49%	0.27%	2.41%
Total			5.42%	100%	11.23%	100%

Table S7. Top 30 most abundant Metazoa subclasses represented in the SSU rRNA identified from the sediment trap metatranscriptomes and organized by percent of total reads in each phylum. Respective metagenomic read percentages are also shown. -, undesignated

Captions for Datasets S1 to S9

Dataset S1. Geochemical chemical measurements of particulate carbon (PC) and particulate nitrogen (PN) fluxes.

Dataset S2. Quality metrics and characteristics of the metagenome-assembled genomes.

Dataset S3. Relative abundance profiles of the dereplicated MAGs. The relative abundance profile of MAGs across the 63 metagenomic timepoints. Taxa that correlated significantly with summer export pulse (SEP) events, elevated carbon flux (ECF) or the Spring-2015 ECF period are indicated.

Dataset S4. Dataset S9. Kendall correlation analyses of eukaryote SSU rRNA taxa associated with the Spring-2015 ECF period.

Dataset S5. Weighted Gene Co-Expression Network Analysis (WGCNA) of prokaryote and eukaryote rRNAs. WGCNA results from rRNA abundance analyses, including *p*-values and module membership values for all modules for each taxon are shown.

Dataset S6. KEGG module completeness of metabolisms of interest in MAGs. Modules analyzed include Photosystem I, Photosystem II, Reductive pentose phosphate cycle (Calvin cycle), Nitrogen fixation, 3-Hydroxypropionate bi-cycle, and Anoxygenic photosystem II.

Dataset S7. Top Genome Taxonomy Database (GTDB) similarities and coverage profiles of nifH genes recovered from the 63 metagenomic timepoints.

Dataset S8. Eukaryote Random Forest classification model output. Complete mean decrease in accuracy list for Eukaryote Random Forest classification model. Columns "SEP" and "non-SEP" represent the normalized variable importance (normalized RF prediction accuracies) for taxa in SEP time periods, versus non-SEP time periods.

Dataset S9. Prokaryote Random Forest classification model output. Complete mean decrease in accuracy list for Prokaryote Random Forest classification model. Columns "SEP" and "non-SEP" represent the normalized variable importance (normalized RF prediction accuracies) for taxa in SEP time periods, versus non-SEP time periods.

Supplementary Information References Cited

1. D. Boeuf *et al.*, Biological composition and microbial dynamics of sinking particulate organic matter at abyssal depths in the oligotrophic open ocean. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 11824-11832 (2019).
2. D. M. Karl, M. J. Church, J. E. Dore, R. M. Letelier, C. Mahaffey, Predictable and efficient carbon sequestration in the North Pacific Ocean supported by symbiotic nitrogen fixation. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 1842 - 1849 (2012).
3. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)* **30**, 2114-2120 (2014).
4. T. Magoc, S. L. Salzberg, FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics (Oxford, England)* **27**, 2957-2963 (2011).
5. B. Bushnell, *BMap: A Fast, Accurate, Splice-Aware Aligner* (United States, 2014).
6. E. P. Nawrocki *et al.*, Rfam 12.0: updates to the RNA families database. *Nucleic acids research* **43**, D130-D137 (2015).
7. E. P. Nawrocki, S. R. Eddy, Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics (Oxford, England)* **29**, 2933-2935 (2013).
8. C. Quast *et al.*, The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590-D596 (2013).
9. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **1303.3997v2 [q-bio.GN]** (2013).
10. J. Oksanen *et al.*, The vegan package. *Community ecology package* **10**, 631-637 (2007).
11. R. C. Team, R: A language and environment for statistical computing. (2013).
12. H. Li. Seqtk: a Fast and Lightweight Tool for Processing FASTA or FASTQ Sequences." (United States, 2013).
13. S. Nurk, D. Meleshko, A. Korobeynikov, P. A. Pevzner. metaSPAdes: a New Versatile Metagenomic Assembler. Genome Research, 27(5), 824-834 (2017).
14. D. D. Kang, J. Froula, R. Egan, Z. Wang, MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
15. D. D. Kang *et al.*, MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
16. Y. W. Wu, B. A. Simmons, S. W. Singer, MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics (Oxford, England)* **32**, 605-607 (2016).
17. C. M. Sieber *et al.*, Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836-843 (2018).
18. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043-1055 (2015).
19. M. R. Olm, C. T. Brown, B. Brooks, J. F. Banfield, dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864-2868 (2017).
20. A. Bankevich *et al.*, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput. Biol.* **19**, 455-477 (2012).

21. P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, D. H. Parks, GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics (Oxford, England)* (2018).
22. T. Seemann, Prokka: rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)* **30**, 2068-2069 (2014).
23. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59-60 (2015).
24. B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder, C. H. Wu, UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics (Oxford, England)* **23**, 1282-1288 (2007).
25. R. L. Tatusov *et al.*, The COG database: an updated version includes eukaryotes. *BMC bioinformatics* **4**, 41 (2003).
26. R. D. Finn *et al.*, The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279-D285 (2016).
27. D. H. Haft *et al.*, TIGRFAMs and genome properties in 2013. *Nucleic Acids Res.* **41**, D387-D395 (2013).
28. T. Aramaki *et al.*, KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics (Oxford, England)* **36**, 2251–2252 (2020).
29. A. Marchler-Bauer, S. H. Bryant, CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.* **32**, W327-W331 (2004).
30. B. J. Woodcroft *et al.*, Genome-centric view of carbon processing in thawing permafrost. *Nature* **560**, 49-54 (2018).
31. P. Virtanen *et al.*, SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261-272 (2020).
32. G. Van Rossum, F. L. Drake Jr., Python reference manual. *Centrum voor Wiskunde en Informatica Amsterdam* (1995).
33. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)* **28**, 3150-3152 (2012).
34. D. H. Parks *et al.*, A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnol.*, 996–1004 (2018).
35. P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
36. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
37. P. Shannon *et al.*, Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498-2504 (2003).
38. L. Breiman, Random forests. *Machine learning* **45**, 5-32 (2001).
39. S. C. Watts, S. C. Ritchie, M. Inouye, K. E. Holt, FastSpar: rapid and scalable correlation estimation for compositional data. *Bioinformatics (Oxford, England)* **35**, 1064-1066 (2019).
40. J. Friedman, E. J. Alm, Inferring correlation networks from genomic survey data. *PLoS Computat. Biol.* **8**, e1002687 (2012).
41. S. W. Olesen, C. Duvallet, E. J. Alm, dbOTU3: A new implementation of distribution-based OTU calling. *PLoS One* **12**, e0176335 (2017).
42. K. Bolar, Interactive Document for Working with Basic Statistical Analysis. *R package version 0.1.0* (2019).

43. R. Kolde, M. R. Kolde, Package ‘pheatmap’. *R Package* **1** (2015).
44. H. Wickham, *ggplot2: elegant graphics for data analysis* (Springer, 2016).