PNAS
www.pnas.org

Supplementary Information for

Biological composition and microbial dynamics of sinking particulate organic matter at abyssal depths in the oligotrophic open ocean

Dominique Boeuf, Bethanie R. Edwards, John M. Eppley, Sarah K. Hu, Kirsten E. Poff, Anna E. Romano, David A. Caron, David M. Karl and E. F. DeLong

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

This PDF file includes:

Supplementary methods and text
SI Appendix Figures S1 to S4
Captions for Datasets S1 to S6
References for SI reference citations

Other supplementary materials for this manuscript include the following:

Datasets S1 to S6

www.pnas.org/cgi/doi/10.1073/pnas.1903080116

Supplementary Information Text
Detailed Methods

Sample collection

From March 15th to November 15th, 2014 two McLane Parflux Mark 7-21 sequencing sediment traps were deployed in parallel at 4000 m (800 m above the seafloor) at Station ALOHA (22° 45' N, 158° W), as previously described (1). For each trap, a total of 21 samples were collected, each sample bottle containing 12 days of accumulated sediment. In the first trap, sinking particles were collected in bottles filled with 250 mL of a filtered, buffered formalin brine solution consisting of filtered surface seawater containing: sodium chloride (5 g/L), sodium borate (1 g/L) and formalin (3% vol/vol final concentration) made up in 4000 mL seawater. The sediment trap was programmed to rotate to a new bottle every 12 days. The samples from this trap were analyzed for particulate carbon, nitrogen, and phosphorus according to the Hawaii Ocean Time-series as described previously (1). In a second trap, sinking particles were collected in bottles filled with 250 mL of nucleic acid preservative solution (5.3 M $(NH_4)_2SO_4$, 25 mM sodium citrate, 20 mM EDTA, pH=5.2) (2) using the same sampling times as formalin fixed traps. Upon sample recovery, the nucleic acid preserved samples were immediately stored at 4°C, and subsequently processed for long term storage by transferring the sedimented biomass using a serological pipet to a 50 ml Falcon tube, and centrifuging at 4,750 rpm at 4°C for 30 minutes in a Beckman Co Allegra X-15R centrifuge. The supernatant was removed, and the pelleted POM was immediately stored at -80°C. Sediment collection cups and Falcon tubes were kept at 4°C during all processing steps.

RNA and DNA extraction and purification

DNA extractions were performed using approximately 70 - 450 mg of frozen sediment sample aliquoted using sterile technique. Extractions and purifications were performed using the MoBio PowerBiofilm DNA Isolation kit (MoBio #2400050, Hilden, Germany) following the manufacturer's protocol. DNA was quantified using Quant-iT PicoGreen® (Invitrogen # P7589, Thermo Fisher, Waltham, MA) kit with a Victor X3 spectrophotometer (Perkin Elmer, Waltham, MA). Genomic DNA size was assessed

using a Fragment Analyzer with DNF- 488 high-sensitivity genomic DNA kit (Agilent, Santa Clara, CA). Average DNA size was 8,968 ±2,338 bp. Samples below 1.67 ng/µl were further concentrated with SC250 EXP Speed Vacuum and UVS800 DDA vacuum system (Thermo Fisher, Waltham, MA). The average DNA yield was 2.3 ng/mg sediment across all samples. RNA extractions were performed using an average of 430 ± 180 mg of frozen sediment sample aliquoted using sterile technique, following mirVana RNA protocol as previously described (3-7). To each sample lysate prior to purification, 25 µl of each internal RNA standards group were added, as previously described (7, 8). A secondary RNA column clean-up step using the MoBio PowerClean Pro RNA Clean-Up Kit (MoBio #1399750, Hilden, Germany) was performed after initial mirVana RNA purification. RNA was quantified using Quant-iT™ PicoGreen® Assay kit (Invitrogen #P11496, Carlsbad, CA) and RNA quality was assessed using DNF-491 high-sensitivity total RNA fragment analyzer kit (Agilent Technologies #491500, Santa Clara, CA). The average RNA yield was 0.21 ng/mg sediment across all samples.

Small subunit rRNA amplicon sequencing and analyses

Bacterial and archaeal rRNA amplicon libraries were prepared by PCR amplification of the V4 hypervariable region of the SSU rRNA using barcoded 515F (5′GTGYCAGCMGCCGCGGTAA3′) and 806R (5′GGACTACNVGGGTWTCTAAT3′) primers as previously described (9), with minor modifications (10, 11)using the recommended PCR thermal profile (12). Amplicon libraries were purified using Agencourt AMPure XP magnetic beads using 1.8X bead ratio to remove primer dimers (Beckman Coulter, Brea, CA). Size was verified with gel electrophoresis and quantification was performed with a Quant-It PicoGreen dsDNA Assay Kit (Thermo Fisher Scientific, Waltham, MA) read on a Victor X3 spectrophotometer. Libraries were normalized and equal concentration pooled, and sequencing was implemented on an Illumina MiSeq sequencer using a V3 reagent kit (2 x 300bp paired end reads) with custom sequencing primers previously described by Caporaso JG. et al. (13). A PhiX sequencing control (10 nM stock) was added to a final representation of 15% of the total sequencing pool. Read trimming was performed using Trimmomatic v0.36 (14). Sequences were then demultiplexed using QIIME2 (9). Error modeling and correction

was performed using DADA2 (9, 15). Finally, taxonomy was assigned using the SILVA v123 database (16).

Eukaryote rRNAs were amplified and sequenced using previously described methods (17). For RNA samples extracted total RNA was reverse transcribed into cDNA (BioRad iScript). The full protocol used to create eukaryotic SSU rRNA metabarcoding libraries can be found online at protocols.io: dx.doi.org/10.17504/protocols.io.hdmb246. Amplicon sequence libraries were created using PCR primers for the V4 hypervariable region of the 18S rRNA gene from the DNA and RNA extracts. The PCR primers were: Forward (5'-CCAGCASCYGCGGTAATTCC-3') and reverse (5'-ACTTTCGTTCTTGATYRA-3'), targeting the V4 hypervariable region (18). V4 amplicons were approximately 400 bp in length and generally provide more phylogenetic resolution and better diversity estimates relative to other shorter hypervariable regions used to study protists (17, 19).

The PCR thermal profile was adapted from Rodriguez-Martinez et al. (19, 20) with an activation step at 98°C for 2 minutes (specific for Q5 High-Fidelity Master Mix, NEB #M0492S, Ipswich, MA), followed by 10 cycles of 98°C for 10 seconds, 53°C for 30 seconds, 72°C for 30 seconds, 15 cycles of 98°C for 10 seconds, 48°C for 30 seconds, 72°C for 30 seconds, and a final extension of 72°C for 2 minutes. PCR products were purified with AMPure magnetic beads (Beckman Coulter #A63881, Brea, CA) and normalized to equal concentrations. Before sequencing samples were multiplexed by performing an additional PCR to anneal Illumina-specific P5 and P7 indices and then combined at equimolar concentrations.

Metagenomic and metatranscriptomic library preparation

A total of 30 ng of genomic DNA was sheared to an average size of 350 bp using a Covaris™ M220 Focused-ultrasonicator™ (Thermo Fisher Scientific #4482277, Waltham, MA) following Illumina's TruSeq Nano Neoprep protocol, with modifications to shear time based on average genomic DNA sizing. 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 quantified using a Quant-iT™ PicoGreen®

dsDNA Assay kit (Invitrogen #P11496, Carlsbad, CA) with a Victor X3 spectrophotometer (Perkin Elmer, Waltham, MA) and library sizes were determined with a Fragment Analyzer using DNF-486 high-sensitivity 35-6000 bp NGS kit (Agilent, Santa Clara, CA). Libraries were sequenced either using a 150 bp paired-end NextSeq500/550 High Output v2 reagent kit (Illumina #FC4042004, San Diego, CA), or a 150 bp paired-end NextSeq500 mid-output v2 reagent kit (Illumina #FC4042003, San Diego, CA), depending on the number of input multiplexed libraries being sequenced.

Quantitative RNA standards were prepared and implemented as described previously (7, 8). Briefly, fourteen RNA standards were generated from DNA templates via T7 RNA polymerase in vitro transcription (IVT) using the MEGAscript High Yield Transcription Kit (Ambion). DNA templates to generate the standards were PCR amplified from *Sulfolobus solfataricus* genomic DNA via PCR amplification and incorporation of the T7 promoter. Prior to RNA purification, 25 µl of each standard group was added to the sample lysate, targeting a final standard concentration of approximately 1% to each sample based on expected total RNA yield.

Metatranscriptomic libraries were prepared with 0.5-50 ng of total RNA using the ScriptSeq v2 RNA-Seq kit (Illumina #SSV21124, San Diego, CA). Unique single-plex barcodes were annealed onto cDNA fragments during the PCR enrichment for Illumina sequencing primers with 12 cycles following manufacture guidelines. Libraries were quantified using Quant-iT PicoGreen® (Invitrogen) with a Victor X3 spectrophotometer and average complementary DNA fragments size was assessed using DNF-486 high-sensitivity NGS fragment analyzer kit (Agilent, Santa Clara, CA). Libraries were normalized to 4 nM final DNA concentration, pooled in equal volumes, and sequenced using an Illumina NextSeq 500 system with a V2 high output 300 cycle reagent kit. The PhiX quality control (Illumina, San Diego, CA) reagent was added to an estimated final contribution of 5% of the total estimated sequence density.

DNA and RNA sequence analyses and workflow

The metatranscriptomic paired-end reads were prepared for analysis with two quality control workflows resulting in 257 million high quality metatranscriptomic and 1.05 billion cleaned metagenomic sequences. The first quality control step screened reads

for primer sequences using Trimmomatic 0.35 (parameters: ILLUMINACLIP:2:40:15) (14) assembled read pairs using PANDAseq 2.9 (parameters: -t .6) (21, 22) and low quality bases removed with Trimmomatic (parameters: LEADING:5 TRAILING:5 MINLEN:45). The second step removed any remaining primer sequences with Cutadapt 1.18 (parameters: -n 3) (23) and low quality bases with Sickle 1.33 (parameters: -l 50) (24).

Using the quality controlled sequence reads, ribosomal RNA (rRNA) sequences were extracted and then split into small and large ribosomal subunit rRNA bins using sortmeRNA 2.0 (using default parameters and all included databases) (14, 24, 25). A total of 80.3 million metatranscriptomic and 5.04 million metagenomic SSU reads were identified.

The taxonomic affiliations of the SSU rRNA sequences from the metatranscriptomic and metagenomic libraries, as well as SSU rRNA amplicon sequences, were assigned by sequence identities to the databases using a 97% similarity threshold. SSU rRNA sequences were compared to the SILVA SSU NR99 database Release 123_1 (16) and the Protist Ribosomal Reference (PR^2) database release 4.5 (26) using Bowtie2 2.2.3 (parameters: --local --sensitive --gbar) (27). Each SSU rRNA read was assigned to a single high quality match using samtools 1.7 (parameters: view -F 2308) (28) to extract primary alignments and filter_blast_m8.py (parameters: -I 98 -L 50) (https://github.com/jmeppley/py-metagenomics) applying an alignment cutoff value of 70 bases and a 97% identity threshold cutoff. Each SSU rRNA sequence was assigned the taxonomy of its most similar match in the database, and counts for each taxon were pooled at different taxonomic levels, to account for the most abundant taxonomic groups identified.

Protein-coding genes and resulting proteins were predicted in cleaned sequences not containing rRNA using Prodigal 2.60 (parameters: -p meta –c) (parameters: -p meta –c) (29). Protein sequences were taxonomically and functionally annotated using homology search by LAST 756 (parameters: -b 1 –x 15 –y 7 –z 25)(30) against the RefSeq database release 75 (31), and functionally annotated by profile search by HMMER 3.1b1 (http://hmmer.org/) against eggNOG database release 4.5(32), respectively.

Contig assessment and quality control of metagenome-assembled genomes (MAGs)

For MAG assemblies, the quality of reads from the 21 metagenomes were filtered using two-passes of BBDuk software as implemented in BBTools 36.32 (http://jgi.doe.gov/data-and-tools/bb-tools/) to remove Illumina adapters, known Illumina artifacts, phiX, and reads with extreme GC values (parameters for 1st pass: "ktrim=r k=23 mink=11 hdist=1 tbo tpe tbo tpe", 2nd pass: "k=27 hdist=1 qtrim=rl trimq=17 cardinality=t 'mingc=0.05 maxgc=0.95"). Sequencing errors were corrected and low abundance k-mers removed using BFC r181 (parameters: "-k 21 -1") (33, 34). Cleaned reads were assembled into 24,553 contigs using MEGAhit (parameters: --presets meta) (34, 35)

Contigs were taxonomically assigned by predicting genes with Prodigal 2.60 (parameters: -p meta –c) (29) and comparing predicted genes to RefSeq database release 79 using LAST 756 (parameters: -F 15 -b 1 –x 15 –y 7 –z 25) (30). The 2,733 assemblies identified as bacterial were imported into Anvi'o 2.1.0 (31, 36) using BWA 0.7.15-r1140 (parameters: mem) (22, 37, 38) to map reads, and using CONCOCT (bundled in Anvi'o) (31, 39) to bin contigs into 25 metagenome-assembled genomes (MAGs). Anvi'o bins were considered taxonomically affiliated if more than 25% of genes within them were annotated with the same taxon. Finally, the taxonomy at the class-level of the entire bin was assessed based on the most abundant taxon among annotations of each bin. In parallel, the taxonomic affiliation and the completion of MAGs have been also assessed by a multiple marker genes approach using CheckM (40).

The average nucleotide identity (ANI) between MAGs, as well as two single-cell amplified genomes (SAGs) previously isolated from traps deployed at 500 m deep (41), were computed using pyANI (42). MAGs were next hierarchically clustered by average-linkeage (UPGMA) of the ANI distance using the R function hclust (34). Based on the length, number of predicted genes, and ANI clustering, 16 bins were selected for further analysis

Deep trap gene catalogue

In addition to the MEGAHIT assembly of all samples, cleaned reads from each sample were also assembled with Meta-3.9.0 (parameters: "--meta -k 21,33,55,77,99,127") (43). A non-redundant gene catalog was built from the 21 individual sample assemblies and the combined megahit assembly by first using Prodigal 2.60 (parameters: "-p meta -c") (29) to predict gene coding sequences on all contigs, then using cd-hit-est V4.7 (parameters: "-c 0.95 -G 0 -aS 0.9 -g 1 -r 1 -d 0") (44) to cluster sequences at 95% similarity over at least 90% of the shorter sequence.

MAG distributions and key-functional genes

Genes and corresponding proteins from MAGs were predicted using Prodigal v.2.6.3 (parameters: -p meta). The abundance and number of transcripts for each MAG gene was retrieved from that of the closest relative in the gene catalogue, as identified by homology search using LAST. The abundance and transcriptional activity of MAG genes was assessed by mapping quality controlled metagenomic and metatranscriptomic reads against them using BWA 0.7.17-r1188 (22, 37). For abundance estimates of each MAG in the metagenomes and metatranscriptomes, the number of metagenomic and metatranscriptomic reads mapped to each MAG's gene were summed and divided by the length of the alignment between the query and the subject. MAGs were annotated by protein sequence homology search against the KEGG database (22) using LAST 756 (30). For the estimation of gene and transcript abundances, the number of reads mapping to genes annotated with the KO number of the function considered were summed and divided by the number of genes with this same KO number.

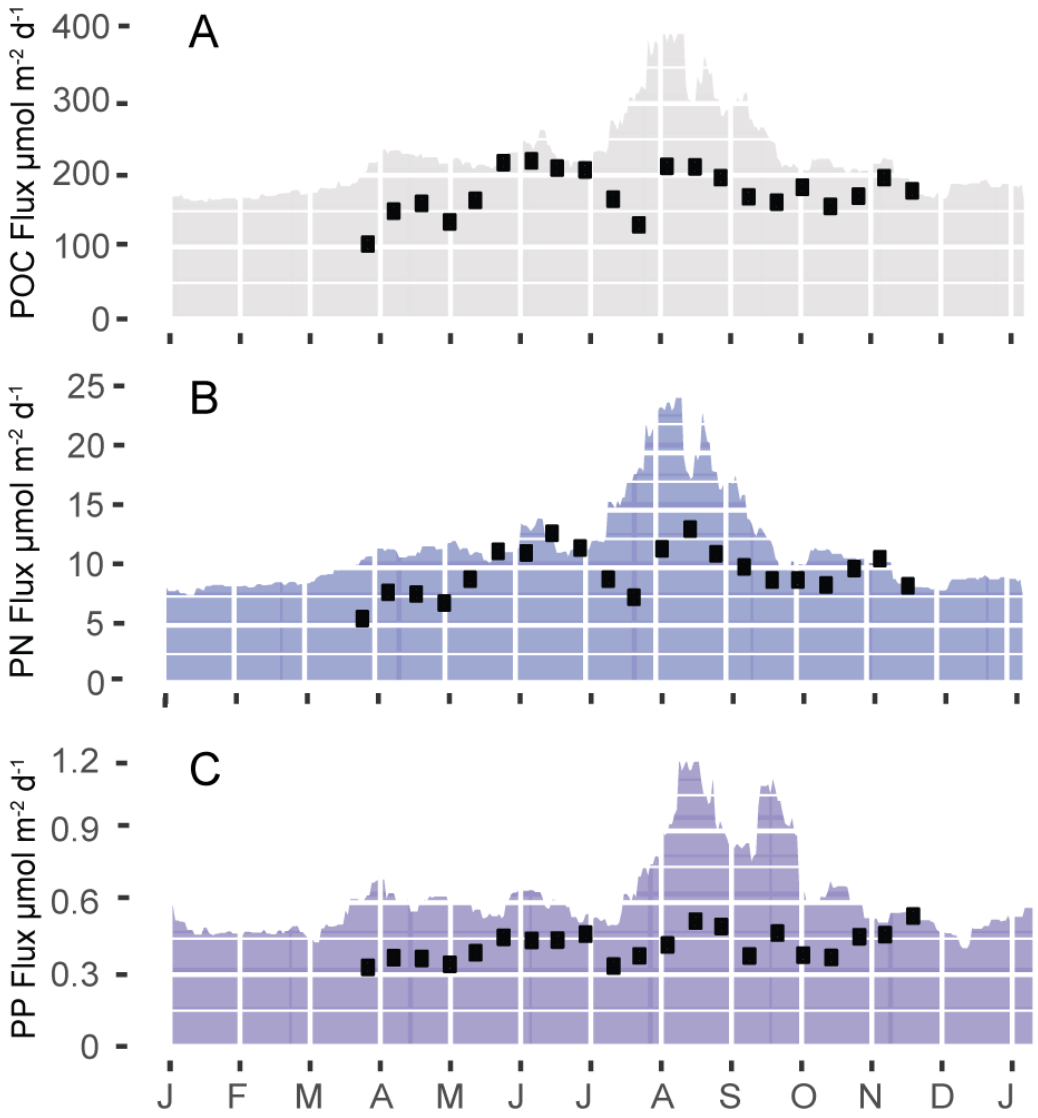

Fig. S1. Formalin fixed sediment trap elemental analyses in 2014, compared to past annual averages. Particulate organic carbon (A), particulate nitrogen (B) and particulate phosphorus (C) flux to 4,000 m at Station ALOHA in 2014 (solid circles), plotted with the climatological averages of POC, PN and PP flux to 4,000 m from 1999 to 2014 updated from the previous data reported by Karl et al, 2012 (1).

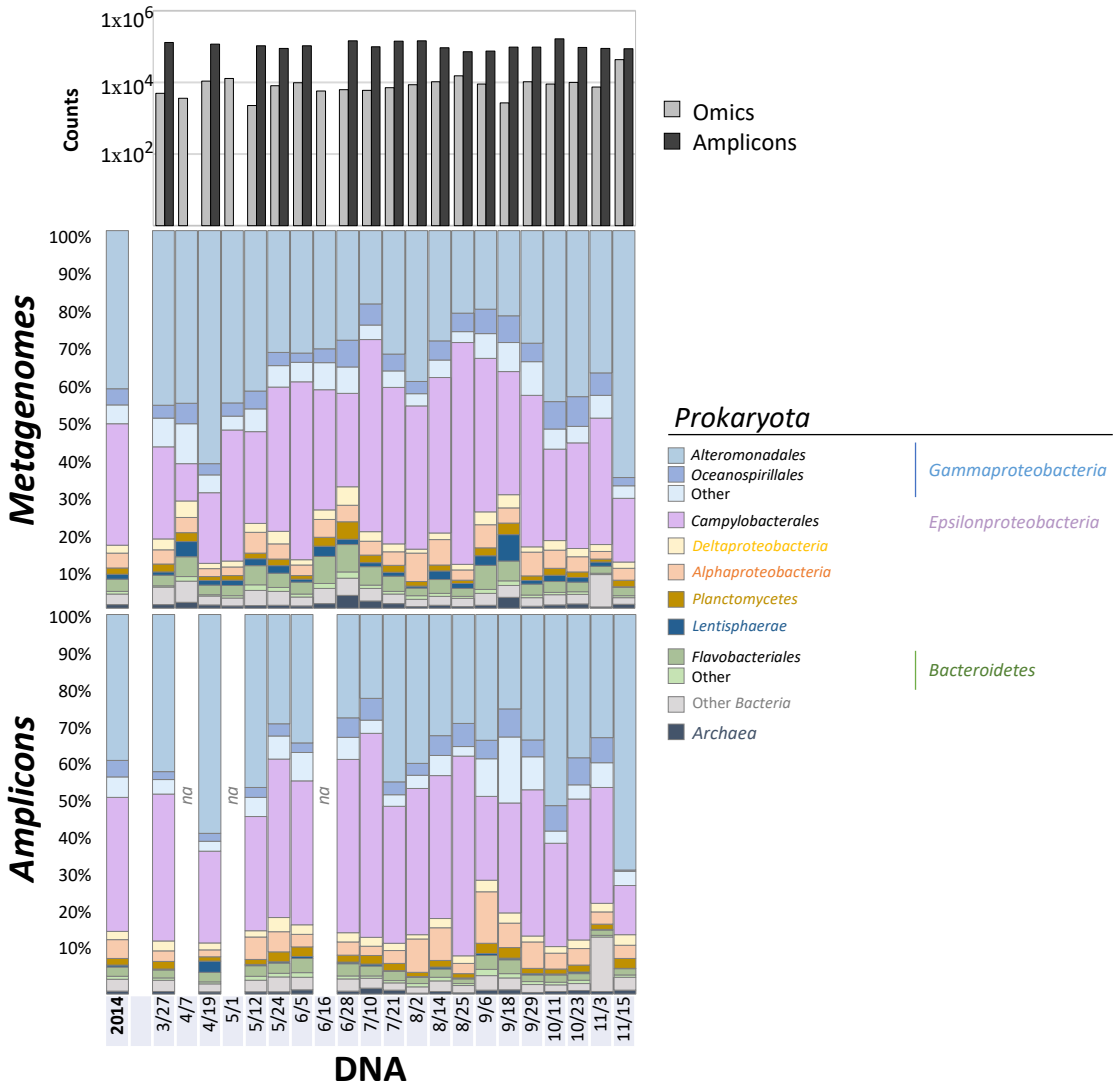

Fig. S2. **Comparison of bacterial SSU rRNA diversity recovered by different methods.** Analyses are derived from amplicon (bottom) or from metagenomic (middle) libraries prepared from particle-associated DNA. The number of counts obtained from each method are displayed by barcharts in the top panel with omics-based counts in grey and amplicon-based counts in black.

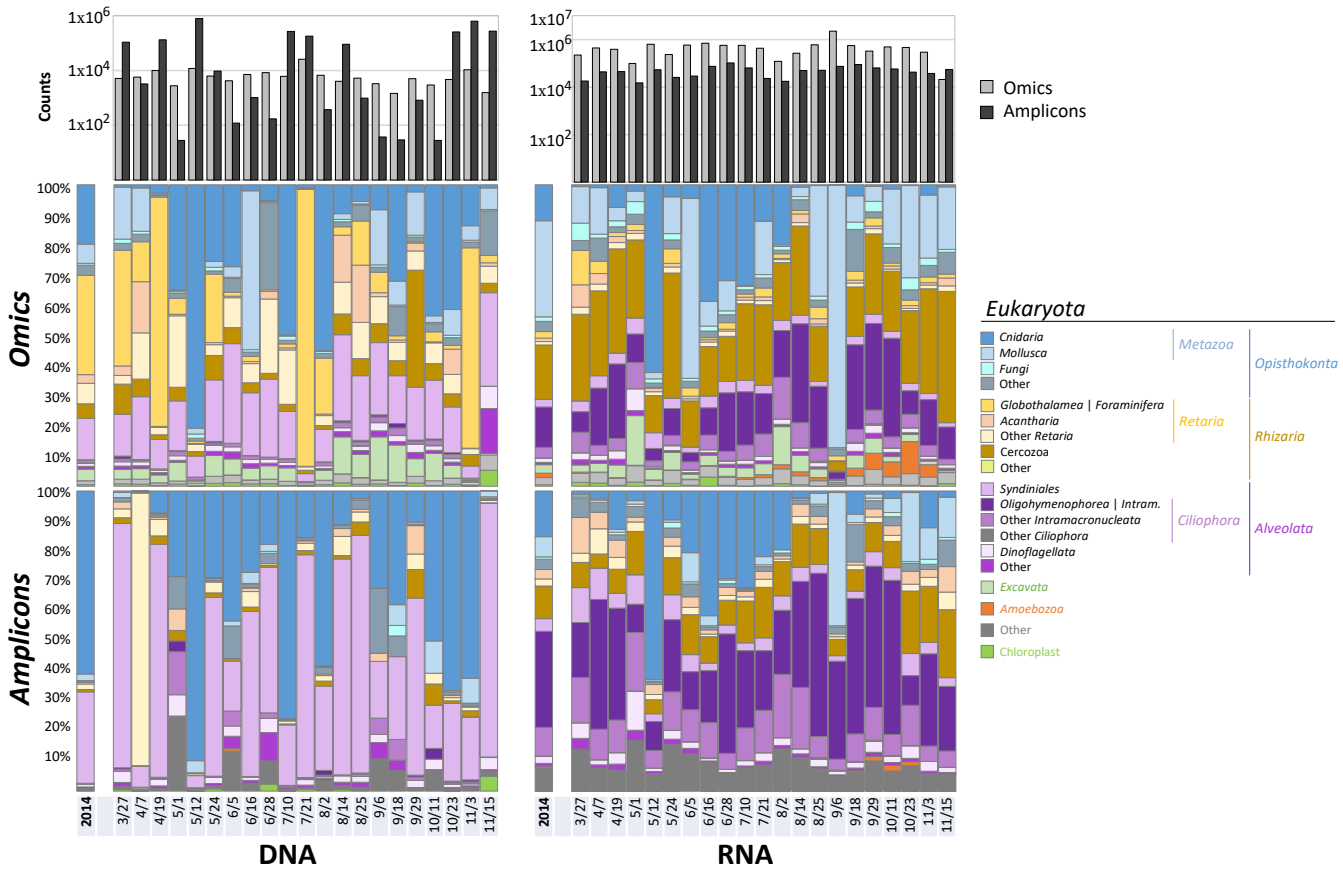

Fig. S3. **Comparison of eukaryotic SSU rRNA diversity recovered by different methods.** Analyses are derived from either amplicon (bottom), or metagenome and metatranscriptome sequencing libraries (middle), prepared from particle-associated DNA (left) or RNA (right). The number of counts ($\log_{10}$) obtained from each method are displayed by barcharts in the top panel, with omics-based (metagenomes or metatranscriptomes) counts in grey and amplicon-based counts in black.

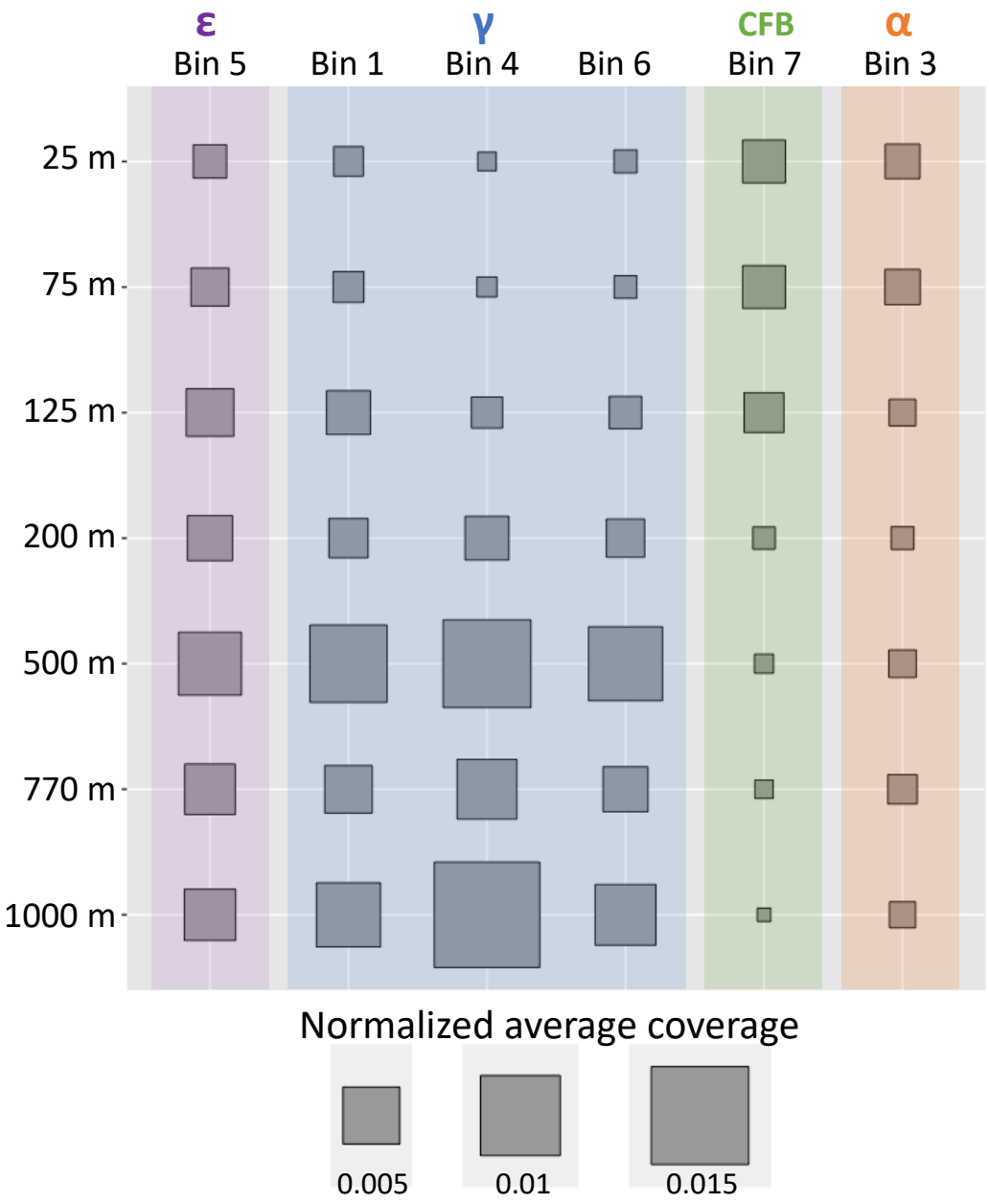

Fig. S4. **Average representation of the deep sediment trap MAGs**

MAGs were recovered in Station ALOHA time-series samples collected from 7 depths (25 m to 1000 m) during January to December, 2011 (45). ε: *Epsilonproteobacteria* (purple); γ: *Gammaproteobacteria* (blue); CFB: *Bacteroidetes* (green); α: *Alphaproteobacteria* (orange).

Bin	Size (Mbp)	Completeness (%)	predicted genes	Annotated (%)	Class	Order	Multiple marker gene affiliation	Most abundant taxon assignation affiliation
Bin_11	0.16	4.2	162	47.5	Epsilonproteobacteria	unresolved	unresolved	Campylobacteraceae
Bin_5	2.45	84.0	2678	62.4	Epsilonproteobacteria	Campylobacterales	*Arcobacter* (genus)	Campylobacteraceae
Bin_9	0.21	-	237	41.1	Epsilonproteobacteria	unresolved	unresolved	Campylobacteraceae
SAG3	0.40	27.6	437	47.0	Epsilonproteobacteria	Campylobacterales	Campylobacterales (order)	n/a
SAG2	0.16	-	161	65.3	Epsilonproteobacteria	unresolved	unresolved	n/a
Bin_12	0.06	10.3	63	81.8	Gammaproteobacteria	Alteromonadales	Alteromonadaceae (family)	Unknown
Bin_8	0.50	15.7	548	79.2	Gammaproteobacteria	Oceanospirillales	Oceanospirillaceae (family)	Oceanospirillaceae
Bin_15	0.08	-	77	54.2	Gammaproteobacteria	unresolved	unresolved	Colwelliaceae
Bin_1	1.90	59.5	1928	62.8	Gammaproteobacteria	unresolved	Gammaproteobacteria (class)	Colwelliaceae
Bin_2	1.60	12.9	1596	60.4	Gammaproteobacteria	Alteromonadales	*Shewanella* (genus)	Shewanellaceae
Bin_4	1.53	13.8	1554	66.2	Gammaproteobacteria	Alteromonadales	Psychromonadaceae (family)	Moritellaceae
Bin_6	1.17	31.0	1080	63.5	Gammaproteobacteria	unresolved	Gammaproteobacteria (class)	Colwelliaceae
Bin_14	0.09	-	83	46.8	Gammaproteobacteria	unresolved	unresolved	Colwelliaceae
Bin_7	0.67	22.7	813	47.9	Flavobacteriia	Flavobacteriales	Flavobacteriaceae (family)	Flavobacteriaceae
Bin_3	1.97	65.3	2139	65.3	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae (family)	Rhodobacteraceae
Bin_16	0.02	-	28	25.0	Alphaproteobacteria	unresolved	unresolved	Unknown

Table S1. Features and affiliation of metagenome assembled genome (MAG) bins recovered from the pooled 2014 sinking POM metagenomic samples.

Additional Datasets

Dataset S1. Taxon counts for all SSU rRNAs in the metagenomic DNA across the 2014 Deep trap time-series: Data_Table_S1_2014_metaGenome_allSSU_lineage_counts

Dataset S2. Taxon counts for all SSU rRNAs in the metatranscriptome RNA across the 2014 Deep trap time-series:
Data_Table_S2_2014_metaTranscriptome_allSSU_lineage_counts

Dataset S3. Taxon counts for all 16S rRNAs in DNA bacterial amplicons across the 2014 Deep trap time-series.
Data_Table_S3_2014_16S_SSU_Amplicon_lineage_counts

Dataset S4. Taxon counts for all 18S rRNAs in DNA eukaryote amplicons across the 2014 Deep trap time-series.
Data_Table_S4_2014_18S_SSU_rDNA_Amplicon_lineage_counts

Dataset S5. Taxon counts for all 18S rRNAs in cDNAs produced from RNA for eukaryote amplicons across the 2014 Deep trap time-series.
Data_Table_S5_2014_18S_SSU_rRNA_Amplicon_lineage_counts

Dataset S6. Collated taxon counts from metagenomes, metatranscriptomes and amplicon data used in main text and SI figures and tables.
Data_Table_S6_Final_SSU_Counts.xlsx

Supplement References

1. Karl DM, Church MJ, Dore JE, Letelier RM, & Mahaffey C (2012) Predictable and efficient carbon sequestration in the North Pacific Ocean supported by symbiotic nitrogen fixation. *Proceedings of the National Academy of Sciences USA* 109:1842-1849.
2. Fontanez KM, Eppley JM, Samo TJ, Karl DM, & DeLong EF (2015) Microbial community structure and function on sinking particles in the North Pacific Subtropical Gyre. *Frontiers in Microbiology* 6:469.
3. Ottesen EA*, et al.* (2013) Pattern and synchrony of gene expression among sympatric marine microbial populations. *Proceedings of the National Academy of Sciences USA* 110:E488-E497.
4. Ottesen EA*, et al.* (2014) Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages. *Science* 345:207-212.
5. Aylward FO*, et al.* (2015) Microbial community transcriptional networks are conserved in three domains at ocean basin scales. *Proceedings of the National Academy of Sciences* 112:5443.
6. Aylward FO*, et al.* (2017) Diel cycling and long-term persistence of viruses in the ocean's euphotic zone. *Proceedings of the National Academy of Sciences*.
7. Wilson ST*, et al.* (2017) Coordinated regulation of growth, activity and transcription in natural populations of the unicellular nitrogen-fixing cyanobacterium Crocosphaera. *Nat Microbiol* 2:17118.
8. Gifford SM, Becker JW, Sosa OA, Repeta DJ, & DeLong EF (2016) Quantitative Transcriptomics Reveals the Growth- and Nutrient-Dependent Response of a Streamlined Marine Methylotroph to Methanol and Naturally Occurring Dissolved Organic Matter. *mBio* 7.
9. Caporaso JG*, et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7:335-336.
10. Apprill A, McNally S, Parsons R, & Weber L (2015) Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquatic Microbial Ecology* 75:129–137.
11. Parada AE, Needham DM, & Fuhrman JA (2016) Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ Microbiol* 18:1403-1414.
12. Gilbert JA*, et al.* (2010) Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project. *Stand Genomic Sci* 3:243-248.
13. Caporaso JG*, et al.* (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A* 108 Suppl 1:4516-4522.
14. Bolger AM, Lohse M, & Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120.
15. Callahan BJ*, et al.* (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* 13:581-583.

16. Quast C*, et al.* (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* 41:D590-596.
17. Hu SK*, et al.* (2015) Estimating protistan diversity using high-throughput sequencing. *J Eukaryot Microbiol* 62:688-693.
18. Stoeck T*, et al.* (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol Ecol* 19 Suppl 1:21-31.
19. Marchant HK*, et al.* (2017) Denitrifying community in coastal sediments performs aerobic and anaerobic respiration simultaneously. *ISME J* 11:1799-1812.
20. Rodriguez-Martinez R, Rocap G, Logares R, Romac S, & Massana R (2012) Low evolutionary diversification in a widespread and abundant uncultured protist (MAST-4). *Mol Biol Evol* 29:1393-1406.
21. Masella AP (2012) PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* 13:31:1471-2105 (Electronic).
22. Kanehisa M, Furumichi M, Tanabe M, Sato Y, & Morishima K (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 45:D353-D361.
23. Martin M (2011) Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. *EMBnet Journal* 17:10-12.
24. Joshi NA & Fass JN (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33)[Software].
25. Kopylova E, Noe L, & Touzet H (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 28:3211-3217.
26. Guillou L*, et al.* (2013) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* 41:D597-604.
27. Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357-359.
28. Li H (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics. *Bioinformatics* 25:2078-2079.
29. Hyatt D*, et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119.
30. Kielbasa SM, Wan R, Sato K, Horton P, & Frith MC (2011) Adaptive seeds tame genomic sequence comparison. *Genome Research* 21:487-493.
31. O'Leary NA*, et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 44:D733-745.
32. Huerta-Cepas J*, et al.* (2015) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* 44:D286-293.
33. Melsted P & Pritchard JK (2011) Efficient counting of k-mers in DNA sequences using a bloom filter. *BMC Bioinformatics* 12:333.
34. R-Core-Team (2017) R: A language and environment for statistical computing, Vienna, Austria. .

35. Li D, Liu CM, Luo R, Sadakane K, & Lam TW (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31:1674-1676.
36. Eren AM*, et al.* (2015) Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 3:e1319.
37. Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754-1760.
38. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* q-bio.
39. Alneberg J*, et al.* (2014) Binning metagenomic contigs by coverage and composition. *Nature Methods* 11:1144.
40. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, & Tyson GW (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* 25:1043-1055.
41. Pelve EA, Fontanez KM, & DeLong EF (2017) Bacterial Succession on Sinking Particles in the Ocean's Interior. *Front Microbiol* 8:2269.
42. Pritchard L, Glover R, Humphris S, Elphinstone J, & Toth I (2016) Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Analytical Methods* 8:12-24.
43. Bankevich A*, et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19:455-477.
44. Li W, Fu L, Niu B, Wu S, & Wooley J (2012) Ultrafast clustering algorithms for metagenomic sequence analysis. *Brief Bioinform* 13:656-668.
45. Mende DR*, et al.* (2017) Environmental drivers of a microbial genomic transition zone in the ocean's interior. *Nature Microbiology* 2:1367-1373.