# Assembled metagenomes collected in the Eastern Tropical North Pacific Oxygen Deficient Zone in April 2012 from R/V Thomas G. Thompson cruise TN278

**Website**: https://www.bco-dmo.org/dataset/733748
**Data Type**: Cruise Results
**Version**: 1
**Version Date**: 2018-04-19

**Project**
» EAGER: Harnessing the power of short-read technology to investigate unexplored microbial communities in the deep euphotic zone (EAGER Deep Prochlorococcus)

| Contributors | Affiliation | Role |
|---|---|---|
| Rocap, Gabrielle | University of Washington (UW) | Principal Investigator, Contact |
| Fuchsman, Clara | University of Washington (UW) | Contact |
| Rauch, Shannon | Woods Hole Oceanographic Institution (WHOI BCO-DMO) | BCO-DMO Data Manager |

**Abstract**
This dataset describes assembled metagenomes collected in the Eastern Tropical North Pacific Oxygen Deficient Zone in April 2012 from R/V Thomas G. Thompson cruise TN278.

## Table of Contents



## Coverage

**Spatial Extent**: **N**:17.043 **E**:-106.543 **S**:16.527 **W**:-107.148
**Temporal Extent**: 2012-04-08 - 2012-04-09

## Dataset Description

Assembled metagenomes collected in the Eastern Tropical North Pacific Oxygen Deficient Zone in April 2012 on cruise TN278.

These data were published in (Fuchsman et al., 2017)

### Methods & Sampling

The GenBank BioProject accession number for this data is PRJNA350692. All samples were obtained from station 136 (106.543W, 17.043N; cast 136) or station BB2 (107.148W, 16.527N; cast 141) on cruise TN278 between April 8-9 2012.

Water from all samples was obtained from Niskin bottles on the CTD Rosette. At station 136 we sampled 10 depths including the oxycline and anoxic zones. Two liters of Niskin water were vacuum filtered onto a 0.2 µm SUPOR filter. At station BB2, a nearby station approximately four liters from 100m, 120m, and 150m were each prefiltered through >30 µm filters and subsequently filtered onto 0.2 µm SUPOR filters. All filters were stored at -80C until further processing on shore. The GenBank BioSample accession numbers are SAMN05944799-SAMN05944812.

DNA was extracted from filters using freeze thaw followed by incubation with lysozyme and proteinase K and phenol/chloroform extraction. A Rubicon THRUPLEX kit was used for library prep using 50 ng of DNA per sample. Four libraries were sequenced on an Illumina HiSeq 2500 in rapid mode (25 million 150 bp paired-end reads per sample) at Michigan State. The other 10 libraries were sequenced on an Illumina HiSeq 2500 in high output mode (40-70 million 125 bp paired-end reads per sample) at the University of Utah. Sequences were quality checked, trimmed, and remaining adapter sequences were removed using Trimmomatic (Bolger et al., 2014). Paired reads that overlapped were combined with Flash (Magoc and Salzberg, 2011). FASTQ files of the raw sequence reads are deposited in the Sequence Read Archive under study SRP092212 and the individual file accession numbers SRS1765745-SRS1765758.

Metagenomic sequences from each sample were assembled independently into larger contigs using two approaches. They have been deposited in the Whole Genome Shotgun archive under the accession numbers:
MOXS00000000
MPLT00000000
MPLU00000000
MPLV00000000
MPLW00000000
MPLX00000000
MPLY00000000
MPLZ00000000
MPLZ00000000
MPMB00000000
MPMC00000000
MPMD00000000
MPME00000000
MPMF00000000

First, for de novo assembly we pre-processed reads with the khmer software package (Crusoe et al. 2015), first using normalize-by-median which implements a Digital normalization algorithm (Brown et al., 2012) to reduce high coverage reads to 20x, followed by filter-abund.py to trim reads of kmers with an abundance below 2, and finally we used filter-below-abund.py to trim kmers with counts above 50 (Zhang et al. 2015). We assembled the khmer processed reads with the VELVET (1.2.10) assembler (Zerbino, 2010), using a kmer size of 45. Velvet assemblies have been released as version 1 in Whole Genome Shotgun archive (accession numbers XXXX01000000).

Second, we also generated a second independent de-novo assembly for each sample using MEGAHIT v1.1.2 (Li et al., 2015) in paired end mode with a minimum contig length of 500. MEGAHIT assemblies have been released as version 2 in the Whole Genome Shotgun archive (accession numbers XXXX02000000).

**Data Processing Description**

BCO-DMO Processing:
-modified parameter names to conform with BCO-DMO naming conventions (replaced spaces with underscores);
-added URLs linking to NCBI;
-split original "Lat Long" column into two separate columns; removed trailing "N" and "W" and removed degrees sign; changed lon to negative values;
-formatted date to yyyy-mm-dd from dd-APR-yyyy;
-removed "m" (units) from Depth column values and "L" from Water_Volume column.

[ table of contents | back to top ]

## Data Files

| File |
| --- |
| **ETNP_ODZ_metagenomes_Apr2012.csv** (Comma Separated Values (.csv), 6.95 KB)<br>MD5:a1f486cf0ae2a52c5ef34a4ea0aaa6a6<br>Primary data file for dataset ID 733748 |

[ table of contents | back to top ]

## Related Publications

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114–2120. doi:10.1093/bioinformatics/btu170
*Methods*

Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B., and Brom, T. H. (2012). A reference-free algorithm for computational normalization of shotgun sequencing data. arXiv:1203.4802v2.
*Methods*

Crusoe, M. R., Alameldin, H. F., Awad, S., Boucher, E., Caldwell, A., Cartwright, R., Charbonneau, A., Constantinides, B., Edvenson, G., Fay, S., Fenton, J., Fenzl, T., Fish, J., Garcia-Gutierrez, L., Garland, P., Gluck, J., González, I., Guermond, S., Guo, J., … Brown, C. T. (2015). The khmer software package: enabling efficient nucleotide sequence analysis. F1000Research, 4, 900. https://doi.org/10.12688/f1000research.6924.1
*Methods*

Fuchsman, C. A., Devol, A. H., Saunders, J. K., McKay, C., & Rocap, G. (2017). Niche Partitioning of the N Cycling Microbial Community of an Offshore Oxygen Deficient Zone. Frontiers in Microbiology, 8. doi:10.3389/fmicb.2017.02384
*Results*

Magoc, T., & Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics, 27(21), 2957–2963. https://doi.org/10.1093/bioinformatics/btr507
*Methods*

Zerbino, D. R. (2010). Using the Velvet de novo Assembler for Short-Read Sequencing Technologies. Current Protocols in Bioinformatics, 31(1). Portico. https://doi.org/10.1002/0471250953.bi1105s31
*Methods*

Zhang, Q., Awad, S., & Brown, C. T. (2015). Crossing the streams: a framework for streaming analysis of short DNA sequencing reads. https://doi.org/10.7287/peerj.preprints.890v1
*Methods*

[ table of contents | back to top ]

## Related Datasets

### IsRelatedTo

University of Washington. marine metagenome Eastern Tropical North Pacific 2012. 2018/01. In: BioProject [Internet]. Bethesda, MD: National Library of Medicine (US), National Center for Biotechnology Information; 2011-. Available from: http://www.ncbi.nlm.nih.gov/bioproject/PRJNA350692. NCBI:BioProject: PRJNA350692.

[ table of contents | back to top ]

## Parameters

| Parameter | Description | Units |
|---|---|---|
| Cruise | Cruise identifier | unitless |
| Station | Station identifier | unitless |
| CTD_cast | CTD cast number | unitless |
| Lat | Latitude; positive values = North | decimal degrees |
| Long | Longitude; positive values = East | decimal degrees |
| Date | Date of sampling | unitless |
| Depth | Depth of sample collection | meters (m) |
| Water_Volume | Water volume collected | liters (L) |
| Sample_Name | Sample identifier | unitless |
| BioProject_Accession | NCBI BioProject accession number | unitless |
| BioSample_Accession_Number | NCBI BioSample accession number | unitless |
| Short_Read_Archive_Accession_Number | NCBI Short Read Archive (SRA) accession number | unitless |
| Whole_Genome_Shotgun_Accession_Number | Whole Genome Shotgun accession number | unitless |
| WGS_Velvet_Assembly | WGS Velvet Assembly | unitless |
| WGS_MEGAHIT_Assembly | WGS MEGAHIT Assembly | unitless |
| Sample_Processing | Description of sample processing | unitless |

[ table of contents | back to top ]

## Instruments

| | |
|---|---|
| **Dataset-specific Instrument Name** | Illumina HiSeq 2500 |
| **Generic Instrument Name** | Automated DNA Sequencer |
| **Dataset-specific Description** | Libraries were sequenced on an Illumina HiSeq 2500. |
| **Generic Instrument Description** | General term for a laboratory instrument used for deciphering the order of bases in a strand of DNA. Sanger sequencers detect fluorescence from different dyes that are used to identify the A, C, G, and T extension reactions. Contemporary or Pyrosequencer methods are based on detecting the activity of DNA polymerase (a DNA synthesizing enzyme) with another chemoluminescent enzyme. Essentially, the method allows sequencing of a single strand of DNA by synthesizing the complementary strand along it, one base pair at a time, and detecting which base was actually added at each step. |

| | |
|---|---|
| **Dataset-specific Instrument Name** | |
| **Generic Instrument Name** | Niskin bottle |
| **Dataset-specific Description** | Water from all samples was obtained from Niskin bottles on the CTD Rosette. |
| **Generic Instrument Description** | A Niskin bottle (a next generation water sampler based on the Nansen bottle) is a cylindrical, non-metallic water collection device with stoppers at both ends. The bottles can be attached individually on a hydrowire or deployed in 12, 24, or 36 bottle Rosette systems mounted on a frame and combined with a CTD. Niskin bottles are used to collect discrete water samples for a range of measurements including pigments, nutrients, plankton, etc. |

[ table of contents | back to top ]

## Deployments

**TN278**

| | |
|---|---|
| **Website** | https://www.bco-dmo.org/deployment/733752 |
| **Platform** | R/V Thomas G. Thompson |
| **Start Date** | 2012-03-17 |
| **End Date** | 2012-04-23 |
| **Description** | Additional cruise data are available from the Rolling Deck to Repository (R2R): http://www.rvdata.us/catalog/TN278 |

[ table of contents | back to top ]

## Project Information

**EAGER: Harnessing the power of short-read technology to investigate unexplored microbial communities in the deep euphotic zone (EAGER Deep Prochlorococcus)**

**Coverage**: South Atlantic, Eastern Tropical North Pacific

*Extracted from the NSF award abstract:*

Recent advances in genome sequencing have transformed marine microbial ecology. Metagenomic datasets have led to the discovery of new protein families, underscored the importance of the rare biosphere and even allowed the study of intra-population diversity of dominant members of the community such as Prochlorococcus. Although open ocean microbial communities have been repeatedly demonstrated to vary significantly with depth, in terms of both phylogenetic diversity and biogeochemical function, the vast majority of sequence available in public databases is from surface waters. Much less is known about microbial communities below the surface of the ocean, even those at 50-200m. Next generation sequencing technologies offer a dramatic reduction in cost per base compared to Sanger or even pyrosequencing approaches. The very short read lengths from these platforms have not yet been successfully applied to mixed communities. This project is a metagenomic study of the microbial communities throughout the water column in the South Atlantic gyre using the ABI SOLiD platform with the following objectives:

- Construct and sequence paired-end metagenomic libraries from the surface, chlorophyll maximum, deep euphotic zone and twilight zone in the South Atlantic gyre using ABI SOLiD technology.

- Apply information theory based software methods to assemble the short read sequences and reconstruct the phylogenetic diversity and metabolic potential of these communities.

- Compare reconstructed sequences to existing metagenomic data from other parts of the world's oceans and to metaproteomic data from the the same location to inform sampling strategies and methodological choices for a proposed global molecular oceanic survey (Biogeotraces).

This project is appropriate as an EAGER because the genomic potential of deep euphotic zone and twilight zone communities will be very different from those in surface waters. Thus, this dataset will reshape our understanding of oligotrophic community function. Second, the massive advance in data generated at a low cost by short-read sequencing platforms has the potential to transform microbial oceanography by allowing metagenomic analyses to be routinely employed in hypothesis-driven experiments. However, new interdisciplinary tools are required to assemble and interpret this data. Until these next generation bioinformatic methods are established, proposals that take advantage of these technological breakthroughs remain high risk.

Broader Impacts: Metagenomic analyses of microbial communities collected on a GEOTRACES compliant cruise in the South Atlantic will provide a pilot dataset demonstrating the potential of short-read sequencing technologies for a future sectional molecular survey program. A post-doctoral researcher and a graduate student will be trained in the interdisciplinary perspectives required to analyze this sequence data and interpret it in a biogeochemical context.

[ table of contents | back to top ]

## Funding


| Funding Source | Award |
|---|---|
| NSF Division of Ocean Sciences (NSF OCE) | OCE-1138368 |


[ table of contents | back to top ]