2020 Ocean Metaproteome Intercomparison Study

Funded by OCB and organized by Mak Saito and Matthew McIlvin
(Woods Hole Oceanographic Institution)
Advisory Committee: Michael Janech (College of Charleston),
Ben Neely (NIST), and Dasha Leary (Naval Research Laboratory)

Version October 30, 2020

Introduction

With increasing interest in the direct measurement of proteins and their biogeochemical functions within the oceans, the metaproteomic datatype is beginning to establish itself as a valuable research and monitoring tool. However, given rapid changes in technology and methods, as well as the overall youth of the metaproteomic field, there is a need to develop confidence in reproducibility and accuracy of these methods. This document describes a proposed effort to conduct the first ocean metaproteomic intercomparison exercise. Following sharing and analysis of samples, a workshop is planned for spring of 2020, where results and methodologies will be discussed with the aim of producing a written product of intercomparison workshop findings.

Methods

Sample Collection and Metadata

Ocean metaproteome filter samples were collected at the Bermuda Atlantic Time Series (31^{o} 40'N 64 ^{o} 10'W) on expedition BATS 348 on June 16th 2018 between 01:00 and 05:00am. Large volume filtration was conducted to produce replicate biomass samples at a single depth in the water column for intercomparisons. All filters subsamples are matched for location, time, and depth. To collect the samples two horizontal McLane pumps were clamped together (Figure 1) and attached at the same depth (80m) with two filter heads units (Mini-MULVS design) on each pump and a flow meter downstream of each filter head. Each filter head contained a 142mm diameter 0.2 μm Supor filter with an upstream 142mm diameter 3.0 μm Supor filter (Figure 2). The pumps were set to run for 240 min at 3 L per min starting at 01:00 local time. Volume filtered were measured by three gauges on each pump, one downstream of each pump head, and one on the total outflow (Table 1). Individual pump head gauges summed to the total gauge for pump 1 (within 1L; 447L and 446.2L), but deviated by 89L on pump 2 (478L and 388.9L). Given that the total gauge is further downstream, we plan to use the pump head gauges.

The pump heads were removed from the McLane pumps immediately upon retrieval, decanted of excess seawater by vacuum, and placed in coolers with frozen blue ice packs, and brought into a fabricated clean room environment aboard the ship. The 0.2 μm filters were cut in eight equivalent pieces and frozen at -80^{o}C in 2 mL cryovials, creating 16 samples per pump that were co-collected temporally and in very close proximity (~1m) to each other that will be used for this study for a total of 32 samples (Figure 3). The 3.0μm filters are not included in this study, but are archived for possible future efforts. The sample naming scheme associated with the different pumps and pumps heads is described in Table 1. Note that pump 1A and 1B samples accidentally had two 3.0 μm filters superimposed above the 0.2 μm filter, and 1B had a small puncture in it, although neither of these seemed to affect biomass on it (presumably the puncture occurred after sampling was completed).

Metagenomic Sequencing and Assembly

A filter slice was extracted and sequenced using a mix of Ion Torrent and Oxford Nanopore sequencing and assembled using SPAdes v. 3.13.1 with Python v. 3.6.8. First, the files were passed to IonTorrent for read-error correction. Then the reads passing QC were assembled with SPAdes. Following metagenome assembly, any contigs smaller than 500 bases were discarded. ORF calling was performed on contigs 500 bases or larger using Prodigal v. 2.6.3 run using metagenomic settings as well as MetaGeneMark by submitting to the MetaGeneMark server (http://exon.gatech.edu/meta_gmhmmp.cgi) using GeneMark.hmm prokaryotic program v. 3.25 on Aug. 11, 2019. ORFs called from both programs were combined and made non-redundant using in-house python scripts that utilize BioPython v. 1.73.

Update Metagenome Assembly & Annotations methods 10/30/20:

The sequencing files were a mix of many short Ion Torrent reads and fewer long reads from Oxford Nanopore. We used the SPAdes v. 3.13.1 pipeline for assembly with Python v. 3.6.8. First, the files were passed to IonTorrent for read-error correction. Then the reads passing QC were assembled with SPAdes. Following metagenome assembly, any contigs smaller than 500 bases were discarded. ORF calling was performed on contigs 500 bases or larger using Prodigal v. 2.6.3 run using metagenomic settings as well as MetaGeneMark by submitting to the MetaGeneMark server (http://exon.gatech.edu/meta_gmhmmp.cgi) using GeneMark.hmm prokaryotic program v. 3.25 on Aug. 11, 2019. ORFs called from both programs were combined and made non-redundant using in-house python scripts that utilize BioPython v. 1.73.

Non-redundant ORFs were annotated using the sequence alignment program DIAMOND (v 0.9.29) with the NCBI nr database (downloaded 12/17/2019). ORFs were also annotated with InterProScan (v 5.29) and with GhostKOALA (submitted to server 1/2/2020). Taxonomy lineages were generated by using the best DIAMOND hit and pulling lineage information from NCBI Taxonomy database using BioPython v. 1.73.

The paired metagenome FASTA file and annotation information is available at:
https://drive.google.com/drive/folders/1j6yxsNbmX1L1DpTEVls8gncU38_LHs4P?usp=sharing

Proposed Intercomparison Study

We will send two ocean metaproteome 0.2 μm filter slice samples, both of which were sampled on the same time and space deployment described above, to 14 laboratories interested in participating in the intercomparison project. While these two samples are intended to be identical, they will be from the different pump heads described above, so total protein abundances will differ slightly due to the volume filtered and can be corrected for in the targeted studies. By providing two samples, it gives each laboratory some opportunity for replication and/or repeated extraction or analytical attempts. The biological diversity within each sample should be close to identical given the sampling approach. We will provide a matching metagenomics database from one of these 0.2 μm subsamples, sequenced by the Naval Research Laboratory, assembled at WHOI, and currently being annotated for peptide to spectrum mapping. In addition we will provide a heterologous overexpressed synthetic protein (QConCat) encoding a suite of isotopically labeled tryptic peptide standards that are representative of this geographic region. Documentation for these standards will be provided in a separate document.

Samples will be shipped starting January of 2020, prioritizing US laboratories conducting oceanographic research, but also including several additional microbiome laboratories to diversify participation, until samples are exhausted. A workshop in Woods Hole will be planned for May 2020 to compare results. An advisory committee consisting of Ben Neely (NIST), Michael Janech (College of Charleston), and Dasha Leary (Naval Research Laboratory) has been assembled and is providing feedback regarding procedures throughout this process.

Specific Intercomparison Efforts:

The intercomparison will include three options for participation, and participants can elect to contribute to any or all of the options as long as they can do so with the provided two samples described above. Option 1 will focus on data dependent acquisition (DDA), given its widespread use in prior metaproteome studies. Option 2 will focus on quantitation using targeted studies, including the use of provided isotopically labeled standards. This effort can be conducted by a variety of means including SRM, MRM, PRM, and DIA approaches. Option 3 will focus on informatic analyses of the DDA datasets from option 1. The intention is to allow a diversity of participants including those only conducting DDA approaches (options 1 and 3), as well as those focusing on informatic components (option 3). In all cases investigators are welcome to conduct analyses with extraction approaches and instrumentation setups of their choosing. Since this is an early intercomparison effort, it is considered to be an opportunity for community building and sharing of knowledge and techniques, as much as it is an opportunity for intercomparisons and data reproducibility assessment. It is assumed from the onset that this effort will lead a future study that will build on these efforts and that will focus on intercalibrations and accuracy and precision more rigorously.

1. Global Metaproteomic Analyses

Each laboratory will conduct total protein extractions according to their preferred methods, followed by data dependent acquisition on their preferred HPLC and mass spectrometry platforms, and data search using their preferred search engine. Guidelines and data on the following parameters is requested for participation in this metaproteome data intercomparison.

1a. Total protein extraction comparison

- Investigators measure total protein using their preferred method after individual lab homogenization.
- Total protein extracted will be calculated by [protein concentration] x [total sample volume] and all values reported.
- Total protein will be normalized to estimated volume of water filtered.
- Metadata: Extraction methods, detergent(s) used, total protein quantitation method used, sample metadata (provided): location, depth, date and time of collection, pore size, liters filtered.

1b. Global Proteome Measurements

- Investigators will prepare trypsin or LysC/trypsin digests and conduct LC/MS/MS using optimal methods and running parameters decided upon per laboratory/investigator. Please avoid labels (iTRAQ, ^{18}O, TMT).

- LC conditions should allow for at least a 60 minute separation. If investigators decide to also conduct two dimensional separations or gas phase fractionations (GPF), submission of 1D analysis is requested to maximize intercomparisons efforts.
- 1D Samples should be injected in triplicate to allow for estimates of variability.
- For 1D analyses: 1 µg per injection is a suggested value. Range of 250ng - 2µg is allowable. It is also recommended to save some material for targeted protein measurements, as additional filter slices will not be available later. No specifications are made for optional 2D or GPF efforts.

Recommended data to be collected:

- Total number of protein Identifications (with recommend false discovery rates $\leq$1%)
- Total number of tryptic peptide identifications (with recommend false discovery rates $\leq$1%)
- Protein attributes (taxon and function; standardized by use of the provided common database)
- Spectral Counts for peptides; Spectral Counts for proteins
- MS1 peak intensities (optional)
- Metadata: Mass spectrometry and chromatographic instrumentation and parameters, Peptide to Spectral Matching software and parameters, genomic database information (provided by study), sample metadata (provided by study; location, depth, date and time of collection, pore size, liters filtered)

2. Targeted Protein Measurements

While it is recognized by the advisory committee that calibrated targeted protein measurements on environmental samples (targeted metaproteomics) represents a relatively new area of research, it was also recognized that this is an area of considerable interest due to its ability to allow spatial and temporal comparisons. To facilitate participation the study will provide an ^{15}N enriched overexpressed protein containing tryptic peptides that are representative of peptides within the metaproteome, as well as reference sequences that correspond to Pierce peptides.

- Investigators will be provided standard protein or semi-tryptic peptides for quantification of targets and its full sequence and precursor mass information.

Recommended data to be collected:

- Creation of external standard curves.
- Concentration of targeted peptides (fmol / L of seawater)
- Limit of Detection for targeted peptides (fmol / L of seawater)
- Limit of Quantification (fmol/L sea water)
- Metadata: Mass spectrometry and chromatographic instrumentation and parameters, targeted peptide sequences for light and heavy peptides (including isotopic composition and position of heavy peptides), targeted protein quantitation software and parameters, sample metadata (location, depth, date and time of collection, pore size, liters filtered per filter, fraction of filter digested)
- Participants should be prepared to share raw file(s) with the group and to deposit them at ProtemXchange to enable any follow up studies the group wishes to pursue during the workshop and to allow publication of results.

3. Informatic Analysis Comparisons

An example raw mass spectra file will be provided at a later point after submission of data from #1, for use in comparison of informatic pipelines. Specifically the file will be a DDA dataset on a similar (but not identical) sample that can be searched using peptide-to-spectrum mapping approaches of the users choice utilizing the provided sample metagenomic database described above.

Recommended data to be collected:

- Total number of protein Identifications (with recommend false discovery rates $\leq$1%)
- Total number of tryptic peptide identifications (with recommend false discovery rates $\leq$1%)
- Protein attributes (taxon and function; standardized by use of the provided common database)
- Spectral Counts for peptides; Spectral Counts for proteins
- MS1 peak intensities (optional)
- Metadata: Mass spectrometry and chromatographic instrumentation and parameters, Peptide to Spectral Matching software and parameters, genomic database information (provided by study), sample metadata (provided by study; location, depth, date and time of collection, pore size, liters filtered)

Target Dates:

Update for Version 10/30/20:

Samples were shipped to laboratories in early 2020. Due to the COVID-19 pandemic during the spring, efforts to follow through on the original timeline were delayed due to laboratory shutdowns and an inability to conduct an in-person workshop. With most laboratories resuming activities at this point in autumn of 2020, we are re-starting the intercalibration effort. *Revised data submission dates are:*

1. Global Metaproteome laboratory datasets by December 1st of 2020
2. Targeted Metaproteome laboratory dataset to be determined (contact us if interested in participating and if we have sufficient participation we will plan a date)
3. Informatic component of the shared RAW file by January 8th, 2021. Prepared raw files for the informatic study will be distributed for informatic analyses on December 2nd 2020 to allow submission of global metaproteome datasets prior to that.

Results will be tabulated anonymized and shared with participants in early 2021 by email and as an initial virtual meeting and discussion. A meeting will be hosted in spring of 2021 (in person if possible, if not virtual) to further discuss the results and prepare a manuscript/report on the findings.

Original Text from Version 1/09/20:

A workshop will be organized in May of 2020. Participating laboratories will receive samples upon request until materials are expended. We recognize that interest may exceed the availability of samples due to the difficulty of obtaining replicate samples for this study, and will remind groups unable to acquire samples due to limited supply that there will be efforts to create a larger study at a later date. The proposed dates for data submission are April 1, 2020 for global metaproteomes and May 1, 2020 for targeted metaproteomics and informatic analyses. Raw files will be distributed for informatic analyses on April 2nd 2020 to allow submission of global metaproteome datasets.

Acknowledgements

Funds for distributing intercomparison samples and a future workshop to discuss results is being supported by Ocean Carbon and Biogeochemistry (OCB). Funds for the research expedition where samples were collected were supported by the US National Science Foundation Biological Oceanography and Chemical Oceanography and the Gordon and Betty Moore Foundation. We thank Sophie Colston, Judson Harvey, and Gary Vora (Naval Research Laboratory) for metagenomic sequencing and Jaci Saunders (WHOI) for assembly and annotation of the paired sample.

Table 1. Volumes filtered through intercomparison sample heads.

Pump / Pump head / sample name	Volume filtered (L)	Volume per 1/8th slice (L)
Pump 2L / BATS 1 / pump 1A	221.6*	27.7
Pump 2R / BATS 2 / pump 1B	167.3*	20.9
Pump 1L / BATS 3 / pump 2A	235.1+	29.4
Pump 1R / BATS 4 / pump 2B	211.1+	26.3

* Pump 1 total gauge = 447L, sum of two pump gauges = 446.2L

+ Pump 2 total gauge = 478L, sum of two gauges = 388.9L, discrepancy of 89L, gauges on pump head are assumed more accurate, as leaks in system could create the additional flow for the total pump gauge.

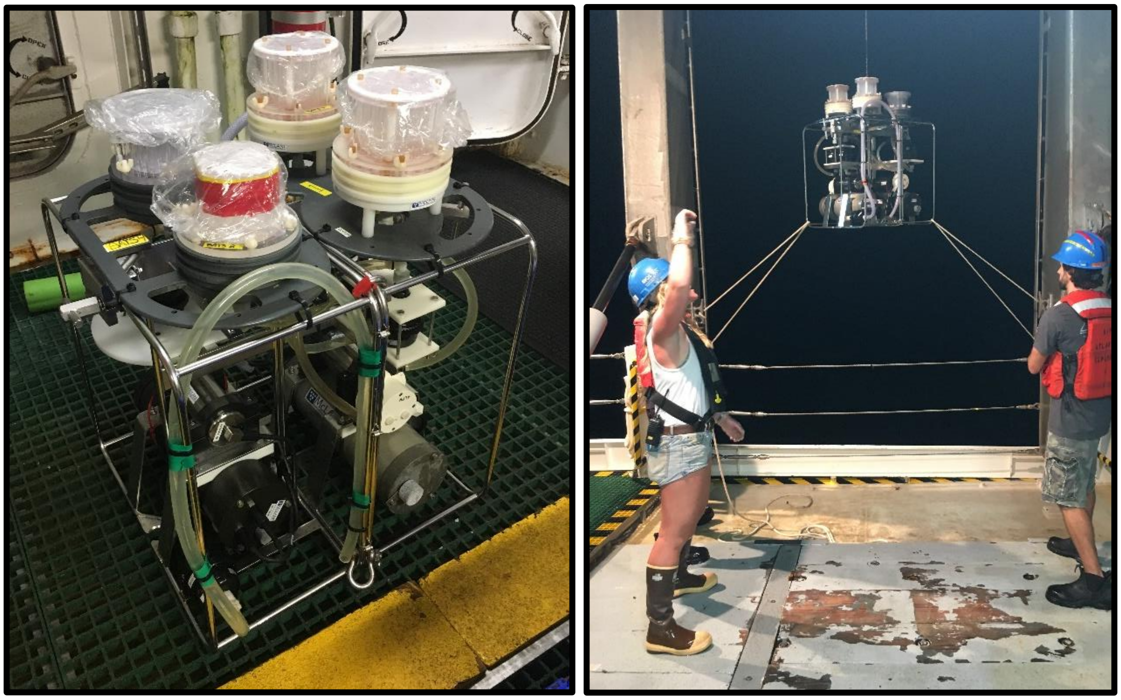

Figure 1. Two McLane pump samplers with two Mini-MULVS sampler heads clamped together and deployed on BATS expedition 348 on June 16th 2018 aboard the R/V Atlantic Explorer.

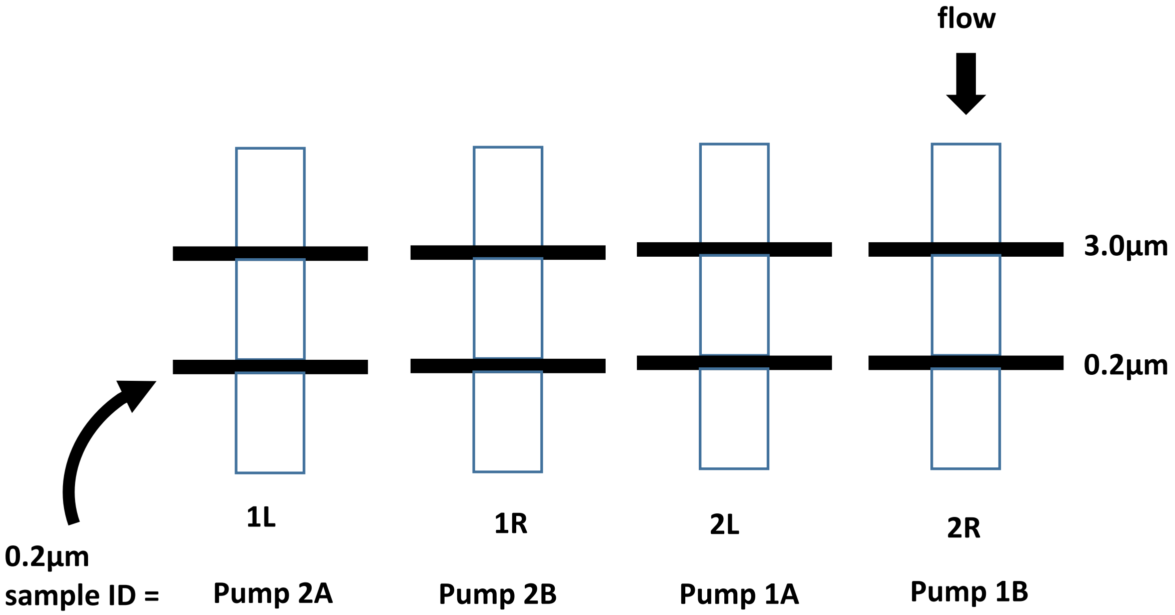

Figure 2. Schematic of sampling, orientation of filters, and labeling schema.

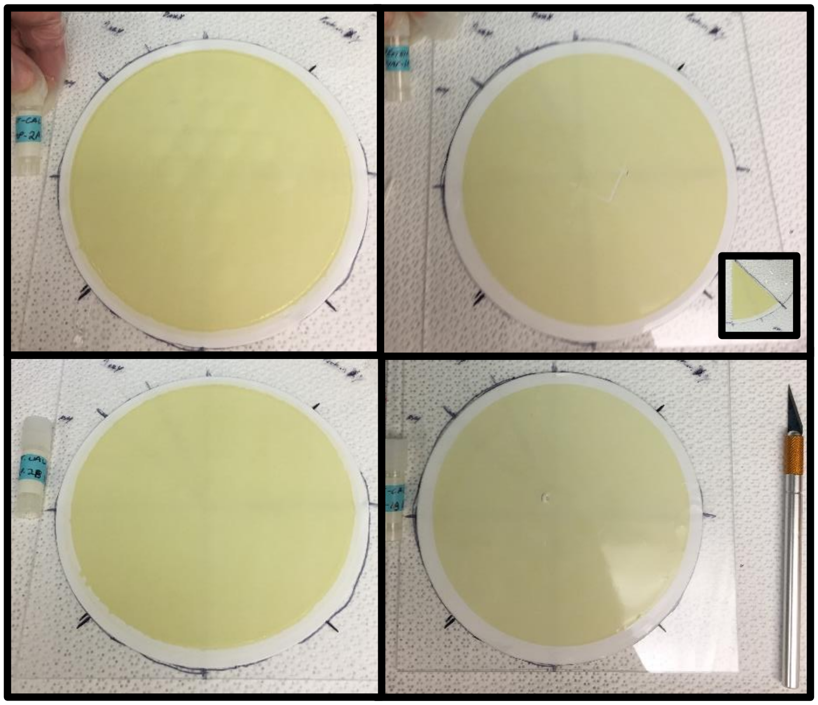

Figure 3. The four 142 mm filter 0.2 μm filters used for this intercomparison study collected by McLane pump (X-Acto knife for scale). Each filter was sliced into 8 fractions (inset) and frozen at -80C in a cryovial. Samples were labeled by pump and pump head (Table 1; pump-2A upper left; pump-1A upper right; pump-2B lower left; pump-1B lower right).

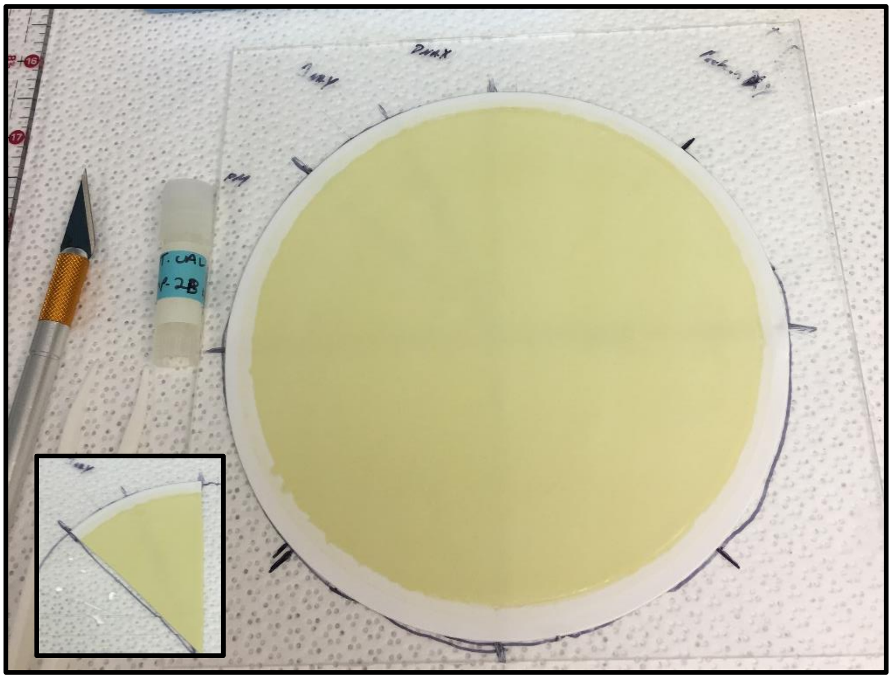

Figure S1. Additional figure of intercomparison samples